Audiovisual Integration in the Recognition of People.

Audiovisuelle Integration beim Erkennen von Personen.

Dissertation

zur Erlangung des akademischen Grades

doctor philosophiae (Dr. phil.)

vorgelegt dem Rat der Fakultät für Sozial- und Verhaltenswissenschaften der

Friedrich-Schiller-Universtität Jena

von MA David Robertson

geboren am 24/12/1981 in Lanark, Schottland.

Gutacher

1. _____

2. _____

_____

Tag des Kolloquiums: _____

TABLE OF CONTENTS

ABSTRACT

Audiovisuelle Integration ist ein wesentlicher Bestandteil der täglichen sozialen Interaktion. Sprache ist deutlich leichter zu verstehen, wenn das Gesicht des Sprechers sichtbar ist. Obwohl Gesichter häufig als verlässliche Informationsquellen betrachtet werden, im Sinne der Identität einer Person, enthält die Stimme eines Sprechers ebenfalls wichtige Informationen über die Person. Ein Grossteil der Forschung zur Personenwahrnehmung konzentriert sich auf unimodale Stimuli, das heisst entweder auf Gesichter oder auf Stimmen allein. Dennoch ist es relativ ungewöhnlich im täglichen Leben ein Gesicht, oder eine Stimme isoliert wahrzunehmen, soziale Erfahrungen sind natürlicherweise auf einer audiovisuellen Ebene. Deswegen ist es wahrscheinlicher, dass uns bekannte Personen im Gedächtnis audiovisuell repräsentiert sind. Mit dem Ziel diese Möglichkeit zu untersuchen, habe ich vier Experimente durchgeführt, um die Effekte der audiovisuellen Integration zu evaluieren. In drei dieser vier Experimente haben Probanden eine Stimmen-Erkennungsaufgabe bearbeitet, in denen verschiedene Gesichterstimuli präsentiert wurden. Im vierten Experiment bearbeiteten Probanden eine Gesichter-Erkennungsaufgabe, bei der Stimmen zusätzlich dargeboten wurden. Die vier Experimente legen den Schluss nahe, dass audiovisuelle Integration ein bedeutsamer Faktor in der Personenwahrnehmung ist. Weiterhin kann die Stärke dieser Effekte, besonders dann wenn bekannte Stimuli involviert sind, die Hypothese unterstützen, dass audiovisuelle Repräsentationen von bekannten Personen im Langzeitgedächtnis existieren.

ABSTRACT

Audiovisual integration is an important part of every-day social interaction. Speech is considerably easier to understand when the face of the speaking person can be seen. Although faces are often seen as a more reliable source of information with regards to a person's identity, voices also hold important information for the recognition of people. Much of the research into person perception has concentrated on unimodal stimuli, concentrating on faces or voices alone. It is however, relatively unusual in normal life to encounter a known face or voice in isolation since social experiences are usually audiovisual in nature. Therefore, it may be the case that the people we know are audiovisually represented in our memory. In order to investigate this possibility, I conducted four experiments to evaluate the effects of audiovisual integration in person perception. In three of the experiments, participants completed a voice recognition task, where various types of face stimuli were presented. In the fourth experiment, participants completed a face recognition task in which different voices were presented. The four experiments strongly suggest that audiovisual integration is a significant factor in person recognition. Furthermore, the strength of the effects observed when familiar stimuli are involved might suggest that audiovisual representations of familiar people exist in long-term memory.

ACKNOWLEDGEMENTS

I would like to thank Prof. Stefan Schweinberger, for his support, understanding and guidance throughout my studies in Germany. I thank Dr. Juergen Kaufmann for his generous assistance with technical issues and Dipl. Psych. Nadine Kloth, for all of her help in collecting stimuli. Many thanks also to Dorit Grundmann for her help in correcting my German grammar in the relevant parts of my dissertation, and for her friendship and support.

I thank also my mother, Jean, father, Robert, and sister, Lynsey, for their love and support throughout studies.

CHAPTER 1:

GENERAL INTRODUCTION

1.1 Multimodal Integration

Everyday experience exposes humans to a multitude of perceptual cues. The senses constantly perceive input from various different sources. The light in a room, the sound of a computer processor's coolant fan, the feel of carpet underfoot, are perceived and interpreted by the brain. These experiences are in most cases, examples of unimodal perception, where they normally exist in isolation and are processed simply based on their respective modalities, visual, haptic and auditory. The different perceptual systems provide a considerable amount of flexibility for perception, so that during sensory deprivation, the other senses can attempt to compensate (eg. In darkness, auditory and haptic perception can compensate for lack of vision to a certain extent (Calvert, Brammer, & Iversen, 1998)). Well-known illusions provide evidence that our perception is not always a true representation of reality. The tendency for perception to show fallibility provides numerous possibilities for researching how perception operates. Classical visual illusions such as "The Hermann Grid Illusion" remain important in investigating the organisation of normal visual perception (Schiller & Carvey, 2005), while auditory (Shepard, 1964) and haptic illusions (Suzuki & Arashida, 1992) can similarly provide insights into sensory perception.

Commonly, however, real-world experiences require the perception of stimuli in more than one modality at the same time. It is likely that the ability to combine unimodal stimuli to form a multimodal percept has evolutionary advantages, since multimodal stimulation in most cases provides a more accurate representation of the world. For early humans, the avoidance of danger was likely to have been greatly improved by the perceptual

integration of signals from more than one modality. The combination of two or more modalities provides the brain with much more information than unimodal inputs can. For example, running one's hand across a brick wall, one perceives the nature of the surface by how it looks, feels and sounds as the hand moves across it, considerably more accurately than if only one sense was available to make the judgement. Research provides evidence for visuo-haptic (Zhou & Fuster, 1997), audio-haptic (Keetels & Vroomen, 2008) and most prevalently, audiovisual integration (AVI). It could be argued that AVI is most important to everyday human interaction, since it affords benefits to object identification (Radeau & Colin, 2001), spatial localisation (Stein & Meredith, 1993), speech (Sumby & Pollack, 1954) and speaker recognition (Rosenblum, Smith, Nichols, Hale, & Lee, 2006).

1.2 Audiovisual Integration

Calvert et al. (1998) suggest that a primary requirement for AVI to occur is that the two modalities should a have a “point of commonality”. Stimuli in different modalities that are presented in spatial and temporal proximity, are often perceived as having a common source (Stein et al., 1993). The perceptual experience is often biased by the modality with the greatest spatial resolution (Calvert et al., 1998), that is, the modality which conveys the most detailed and reliable information about the experience. In the context of visual and auditory stimuli presented in close spatio-temporal proximity, the visual stimuli often biases the perception of the location of the auditory stimulus. This is generally known as the ventriloquist effect (Howard & Templeton, 1966) and despite the implication that it is a phenomenon related directly to speech, integration of simple visual and auditory stimuli (such as a light-flash and a beep-tone) can be achieved. Specifically, the effect relates to temporally proximal, but spatially disparate stimuli, which are perceived as emanating from a common source. When a light-flash is presented simultaneously with a tone which is presented from a different position, the stimuli are perceived as being closer to each other, or coming from the same source.

Single-neuron studies in animals have found that neurons in the superior colliculus (SC) respond superadditively when such audiovisual stimuli are presented in spatiotemporal proximity, while the activation of these neurons are inhibited by the presentation of unimodal, or asynchronous stimuli (Stein et al., 1993). Absolute spatiotemporal synchrony is not a requirement of these effects, as might be clear from the spatial disparity of the visual and auditory sources inherent in the ventriloquist effect. A spatiotemporal window

for this effect, that is, the window in which the two modalities are most often perceived as a single event, has been suggested to range from ±3º of spatial disparity, and 100 milliseconds of temporal disparity (Lewald, Ehrenstein, & Guski, 2001). It has also been shown that the perception of synchrony of audiovisual stimulus is more likely when the auditory stimulus lags the visual stimuli within this time window (Lewald & Guski, 2003), which may reflect the calibration of multisensory perception in order to deal with the natural asynchronies caused by the physical properties of light and sound. Varying certain attributes of the spatiotemporal relationship between the stimuli can have illusory effects. For example when a single light-flash is presented with multiple beep-tones, the light-flash can be perceived as multiple flashes (Shams, Kamitani, & Shimojo, 2002). Furthermore, when the intensity of the visual stimulus is varied, the auditory stimulus has been shown to cause an increase in the perceived intensity of the visual stimulus (Stein & Wallace, 1996).

Spatiotemporal proximity need not be absolute, but it is more important for simple multimodal stimulus pairings than for more complex stimuli. It is suggested (Calvert et al., 1998), that shared information-content is an important factor to the robustness of the integration of two stimuli. Since the stimuli in simple multimodal audiovisual pairings (eg. A light flash and beep-tone) do not share any specific attributes apart from their similar onsets, the integrated percept is relatively fragile. The finding of a ±3º window of spatial disparity (Lewald et al., 2003), in which the two stimuli are perceived as the same event, is small in comparison to the spatial tolerance seen for more complex stimuli. Furthermore, the 100ms temporal window suggested by Lewald et al. (2003) is also relatively small compared to more complex stimuli, such as speech. Therefore, for stimuli

with low informational content, and few shared attributes, it is less likely that they will be integrated. The perception of such stimuli as belonging to the same event is therefore dependent on a high degree of spatiotemporal proximity.

Stimuli that contain more information-content and shared attributes, such as speech stimuli, are subject to less stringent cognitive constraints. It has been suggested that more complex stimuli causes a reduction of the dependence on the spatial and temporal aspects of the presented stimuli (Jones & Jarick, 2006). Audiovisual speech stimuli are able to be integrated with spatial disparities of up to ±38º (Calvert et al., 1998), and with temporal disparities of up to 180ms (Munhall & Vatikiotis-Bateson, 2004). Although these cognitive constraints are less strict than those for simple stimuli, a certain degree of spatiotemporal proximity is still highly important to AVI for complex stimuli. The shared information between auditory and visual speech stimuli – such as the temporal frequency and amplitude of an utterance (Summerfield, 1992) – means that when the informational attributes match across modalities, they are naturally assumed to belong to the same underlying event, making spatiotemporal disparities slightly less influential, making integration of the two modalities more likely.

The aforementioned requirements form the basic tenets of what is considered the "cognitive compellingness" of an audiovisual pairing (Warren, Welch, & Mccarthy, 1981). It is suggested that the cognitive compellingness of a stimulus-pairing governs whether it shall be perceived as a single event, or as two separate auditory and visual events. Low cognitive compellingness would suggest that the modalities do not share any attributes and are more sensitive to spatial and temporal disparities. An example of this may be

when the sound of a speaking voice is presented from speakers at a different location to a static visual stimulus, such as a black rectangle. The stimuli share no attributes, so the subject would not be compelled to perceive the two stimuli as part of the same event. High cognitive compellingness, on the other hand, concerns stimuli that share complex attributes and are also presented in close spatial and temporal proximity. A relevant example is the perception of natural speech, where the tonal attributes of the voice can be closely correlated with the amplitude and temporal characteristics of the face. When in synchrony, and the modalities are perceived to be from a similar location, participants would be strongly compelled to perceive the visual and auditory stimuli to originate from the same unitary event, rather than being a combination of independent voice and face stimuli.

1.3 Audiovisual Integration in the Perception of Speech

During interpersonal communication in real-life situations a person's facial articulatory movements are typically observed at the same time as their voice is heard. Typically the audiovisual input serves to make it easier to perceive what a person is saying, particularly if one of the modalities is disrupted in some way. For instance, in a noisy environment, it has been found that viewing a speaker's face can result in a strong enhancement in the ability to perceive what the person is saying (Sumby and Pollack, 1954; Grant and Seitz, 2000; Ross, Saint-Amour, Leavitt, Javitt and Foxe, 2007). This highly efficient system of multimodal interaction results in significant benefits when a single modality is difficult to perceive, but this efficiency can also be exploited by unusual situations. The assumption of unity when auditory and visual events simultaneously occur, can be manipulated to create audiovisual illusions. As alluded to previously, a classic example of such effects can be witnessed in the "ventriloquist illusion" (Howard et al., 1966), where a sound source is perceived as coming from the same spatial location of approximately time-synchronized visual motion, although the sound is in fact generated by a different source at a slightly different location. More specifically, to use the example of how the phenomenon got its name, the ventriloquist speaks without moving his lips, while his puppet's mouth moves in approximate synchrony with the heard speech. In the absence of another possible perceptual source of the heard voice, the movements of the puppet's mouth and the heard voice are bound as a unitary event. A second classic audiovisual illusion is the "McGurk effect" (McGurk and Macdonald, 1976). This audiovisual illusion pertains to the finding that in the majority of cases, when viewers are presented with an auditory syllable (e.g. /ba/), synchronised with a face articulating an incongruent visual

syllable (e.g. /ga/), they often report hearing an entirely new syllable (/da/). Here, the plosive /ba/ syllable is presented auditorily, while the glottal /ga/ syllable is seen. The ambiguity of the visual /ga/ in presentation with an unexpected auditory stimulus, causes AVI processing to alter auditory perception towards the most likely result. Since the visual stimulus does not look like a plosive (where the lips would be together in producing the consonant), the assumed unity of the event suggests that what was heard came from the face presented, so the AVI compromise is the perception, or "hearing" of /da/. Whereas much of the classical AVI literature was concerned with other aspects of AVI (Howard et al., 1966; Sumby et al., 1954), the McGurk illusion represents a relatively rare, but well-researched, example of AVI for stimulus identification (Calvert et al., 1998).

Apparent unity therefore, does not always result in an enhancement of recognition, but can be manufactured to elicit illusory percepts as in the McGurk illusion. This effect has been shown to be robust, so that even in cases where the face and voice are of different gender, the strength of the McGurk illusion is not affected (Green, Kuhl, Meltzoff & Stevens, 1991). As a qualification, another study found the strength of the McGurk illusion to be reduced when *familiar* faces and voices of different speakers were combined, suggesting that AVI in *speech* perception may not necessarily be independent of *speaker* recognition (Walker, Bruce & O'Malley, 1995). The McGurk effect is traditionally considered to be relatively independent of voluntary control, as the illusion remains robust even when participants are informed of the effect (van Wassenhove, Grant, & Poeppel, 2005), but see (Soto-Faraco & Alsius, 2007), for a qualification).

Approximate time-synchronisation of visual and auditory stimuli is important to achieving AVI effects, and synchronisation is often a significant contributor to a percept of "unity".

However, it has been suggested that perfect synchrony of the stimuli in both modalities is not crucial to the perception of the McGurk effect. Research manipulating asynchrony to test its influence on the McGurk effect (Munhall, Gribble, Sacco, & Ward, 1996; van Wassenhove, Grant, & Poeppel, 2007), suggest that there is a small time-window for integration, within which the McGurk illusion is most likely to be perceived. The research into the time-window of integration for the McGurk effect found that there was a greater tolerance for asynchronous presentation when the auditory stimulus lagged behind the visual stimulus in comparison to the auditory stimulus leading the visual stimulus. In fact, both studies found that where the auditory lag the visual stimuli slightly, the McGurk illusion, and hence AVI, occurred more often (similar to the “simple” AVI effects reported by Lewald et al., (2003)). Audiovisual processing may be predisposed to tolerate such slight asynchronies due to the differing velocities of sound (330 m/s) and light (approximately 3×10^8 m/s). The synchrony of the stimuli in the two modalities would vary depending on the distance between the observer and the stimulus, meaning that the same audiovisual event would stimulate the sensory organs with a certain degree of time offset.

1.4 Audiovisual Integration in the Perception of Identity

Face to face communication in normal situations is essentially an audiovisual experience. However, face recognition and voice recognition research have usually been separated, focusing on either visual or auditory processing only (Bruce, 1990; VanLancker, Kreiman, & Emmorey, 1984). Such research shows that it is of course possible to recognise familiar people from their faces and voices alone. Nevertheless, audiovisual information may have a significant role to play in person identification. While there has been a significant amount of research in face-voice AVI recently (Maravita, Bolognini, Bricolo, Marzi, & Savazzi, 2008; Zekveld, Kramer, Vlaming, & Houtgast, 2008; Bernstein, Auer, Wagner, & Ponton, 2008), the majority of this research is in speech perception, and very few studies have directly addressed the potential role of AVI in person recognition (Campanella & Belin, 2007).

A certain amount of relevant research has been done into the role of audiovisual integration in the perception of person identity. There exists evidence that face identity can be primed by voice identity. Ellis, Jones & Mosdell, (1997), have shown that over short time-intervals, crossmodal priming occurs. They demonstrated that the presentation of a familiar voice-prime followed immediately by a face of *corresponding* identity, resulted in a significant improvement in performance. Similar results were demonstrated for face primes in relation to voice test stimuli. Familiar face-voice priming has been shown to occur even with long intervals (10 minutes) between prime and target (Schweinberger, Herholz, and Stief, 1997).

Sheffert & Olson, (2004), conducted an experiment into voice-learning and word recognition. The experiment consisted of a “familiarisation” phase, in which participants first became familiar with the voices of a number of speakers, followed by a “talker-training” phase, in which they were required to identify the speaker presented. At the end of each trial during these initial phases, the name of the correct speaker was indicated by the experimenter. The important difference between participants was that some were assigned to an auditory-only training condition, while other participants were assigned to an audiovisual training condition. After the training phases, participants underwent a “generalisation” phase in which the same speakers were presented, but uttering a new set of words. This phase was auditory-only, regardless of the modality of the training phases completed by the participants. Thereafter, a word recognition test was conducted to study the effects of speaker familiarity on word recognition memory. The findings of this study showed that voice learning and recognition was greatly improved by audiovisual training phases compared to the auditory-only condition. (2004) suggest that the additional visual information about the speaker’s idiosyncratic speaking style is compatible with the speaker’s auditory attributes, and may therefore lead to better encoding of voice identity. Furthermore, they found that word recognition was better when words were spoken by familiar speakers compared to words spoken by unfamiliar speakers, which might suggest that speaker and linguistic perception are intertwined.

Idiosyncratic facial speech patterns have also been shown to be of relevance to unfamiliar speaker perception. Kamachi, Hill, Lander, & Vatikiotis-Bateson, (2003) sequentially presented unfamiliar dynamic faces and unfamiliar voices to participants. They found that

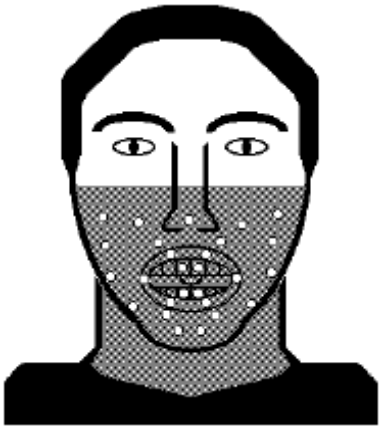


Figure 1 A schematic example of the point-light configuration used by (Rosenblum & Saldana, 1996)

even though the auditory and visual stimuli consisted of different sentences, voices could be matched to faces, and faces could be matched to voices at above-chance levels. Similar to Sheffert et al., (2004), they suggested that the existence of bimodally available dynamic information about speaker characteristics allowed the faces and voices to be matched despite the participants being completely unfamiliar with the speakers. Furthermore, (Rosenblum et al., 2006) reported above-chance face to voice matching even when only dynamic facial information was presented. Using a point-light technique (see Figure 1), where illuminated spots were visible on a face in complete darkness, they were able to isolate facial speech movements. They compared the normal and idiosyncratic speech movements with conditions in which the movements were distorted. They found that face-voice matching was significantly better for the conditions in which the normal facial movements were presented. Rosenblum et al., (2006) highlight particularly clearly the importance of isolated facial movements to the relationship between a speaker's face and voice. Perhaps an auditory analogue to point-light facial movement displays, are voices that have been transformed in the temporal domain (Lachs & Pisoni, 2004a; Lachs & Pisoni, 2004b). In these two studies, it was found that

face-voice matching is still achievable even when the modalities are significantly degraded.

In the case of familiar people, it seems conceivable that multimodal representations of a familiar person's identity may be encoded in long term memory. Such crossmodal effects suggest that dynamic representations of familiar people may exist in long term memory. My research group recently provided the first direct evidence that AVI occurs in the recognition of familiar voices (Schweinberger, Robertson, & Kaufmann, 2007). That study found that dynamic face stimuli had the largest influence on the recognition of a voice as familiar of unfamiliar. When the face was of *corresponding* identity to the voice, response-time and accuracy was generally faster and more accurate compared to when static pictures were presented. When *noncorresponding* dynamic stimuli were presented, response-time and accuracy were generally slower and less accurate, while *noncorresponding* static stimuli had very little effect on voice recognition performance. The following chapters attempt to extend these results.

1.5 General Overview of Methods

The four experiments used similar designs and, with some alterations in Experiment 3 and 4, the same stimuli. In order to ensure that synchrony was the same across all stimuli, the visual and auditory clips were standardised, so that each could be combined with any other without systematic benefits or costs to synchrony regardless of the stimulus identity.

Video-Editing

Videos were standardised according to an average of the consonant onsets (CO) contained in the original clips. COs were identified in Adobe Premiere Pro 1.5 by moving frame by frame through the clip until the first frame at which the CO could be heard in the video's audio track was localised. The CO frames were noted for each word of the sentence: ***D***u ***b***ist ***d***och (***w***)as ***d***u ***d***enkst ("You are what you think"). This sentence was chosen due to the number of plosives. Plosives, also referred to as "stop-consonants", are produced by stopping the airflow in the vocal tract. Due to the nature of plosives, they are clearly identifiable in an auditory clip, in comparison with other speech sounds, since they have a clearly defined onset. Initial plosives were used as the main time-markers in the utterance, although the fricative /v/ was also used, although the precision of identifying fricatives is less reliable than with plosives. Once all of the COs were identified for each speaker, the precise timings of the COs were employed in calculating the average COs for each word across all of the speakers (see Table 1).

Consonant Onset (CO) averages for standardising video stimuli						
CO	***D***u	***b***ist	***d***och	***w***as	***d***u	***d***enkst
Video *frames* (ms)	*6* (240)	*10* (400)	*21* (840)	*36* (1440)	*45* (1800)	*49* (1960)
Vowel Onset (VO) averages for standardising audio stimuli						
VO	D***u***	b***i***st	d***o***ch	w***a***s	d***u***	d***e***nkst
Audio (ms)	306	425	900	1515	1865	2023

Table 1 The consonant-onsets (CO) for each video, and the vowel-onsets (VO for each audio clip, averaged across all speakers. Frame consonant onsets are displayed for the videos, with the CO timing in ms displayed in parenthesis (1 frame = 40ms).

Thereafter, all videos were edited to match this average so that the frame at which each CO occurred was identical for each speaker. This involved duplicating or deleting frames where there was relatively little movement of the face, such as between words, where deleted frames are less noticeable – or at the end of non-rounded (or elongated) syllables, where extra frames are also less noticeable.

Audio-Editing

In Adobe Audition, the auditory tracks were opened from the previously edited video files described above and the exact time-points at which each *vowel* was voiced, following a consonant, were identified using the relatively fine-grained analysis afforded by Adobe Audition (Figure 2*a*). Vowels following plosives are clearly identified as the wave transforms to a recognizable, proximal periodicity as the vowel is voiced. The vowel-onsets (VO) in the video-audio were noted following each plosive as well as for the single fricative. Since the audio tracks came from the edited videos, VO differences between those audio-from-video clips were never more than 40 ms, that is, within one frame of a difference. This allowed the precise localisation of the vowel onsets for each video track.

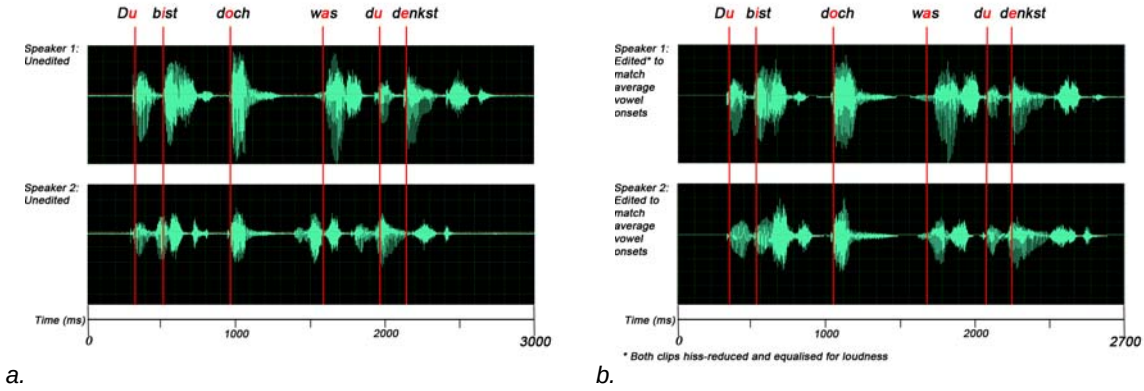


Figure 2a,b An example of two auditory clips having been edited to baseline. *2a* displays the auditory clips prior to editing, and *2b* shows the auditory clips after editing to the standardised timing (averaged across all speakers).

Average VOs across all the stimuli were again calculated (see Table 1, above). The auditory clips which were separately recorded from the microphone (as opposed to the clips of lesser quality from the camera microphone, which were used only for temporal localisation purposes) were then edited to match the calculated average VOs so that they were identical across all audio stimuli (Figure 2*b*). This allowed the combination of any speaker's video and audio stimulus with virtually perfect synchronisation.

In order to evaluate the possible audiovisual integration effects in each experiment, the mean reaction-times and percentage-correct data were analysed for each condition. By demonstrating benefits and costs of audiovisual conditions in relation to the unimodal baseline, it was expected that such a general analysis strategy would most clearly evaluate the effects of audiovisual presentations compared to when only one modality was available. As a previous study, similar to Experiment 1 (Schweinberger et al., 2007) had suggested that benefits and costs to voice recognition performance are conferred by dynamic audiovisual stimuli compared to voice-only stimuli, analysing the mean

response-times and percentage-correct data was also important in extending the findings of that study. Analyses of variance were performed on all data, to ascertain the significance of any costs and benefits in the conditions. Each condition was directly compared to the unimodal condition, to more clearly demonstrate the magnitude of any AVI effects. Finally, dependent on significant interactions between presentation condition and familiarity, difference-scores were calculated by subtracting the voice-only baseline from each audiovisual condition in order to further clarify the respective magnitudes of the AVI effects.

CHAPTER 2:

AUDIOVISUAL INTEGRATION DURING VOICE RECOGNITION

EXPERIMENT 1

Introduction

With perhaps the exceptions of telephone conversation, and more recently e-mail, human social interaction is most often audiovisual in nature. As alluded to in Chapter 1, my research group provided the first direct evidence of AVI in familiar voice recognition (Schweinberger et al., 2007). In that study, it was found that benefits in voice recognition (relative to a voice-only condition) occurred when a voice was presented together with a face of matching identity. Costs to voice recognition, when a voice was presented with a face of a different identity were also found. Importantly, these effects were much stronger when the face was dynamic – moving in a synchronised manner with the voice – compared to when a static picture of the face was shown. The finding of much stronger effects for dynamic (that is, synchronised motion) videos than for static pictures is also in line with the idea that synchronisation is important in leading to cognitive compellingness and a perception of unity (Warren et al., 1981). Moreover, this consistent pattern of benefits and costs was seen only when the voices were familiar. It should be noted that there may be constraints linking individual faces and voices even for unfamiliar speakers, such as vocal tract and body size (Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003). However, our previous findings were interpreted as suggesting that multimodal representations of familiar people are stored in long-term memory, and that links between a specific known face and the *corresponding* voice exists beyond such constraints.

When considering these findings in the context of a potential role of dynamic visual information in person recognition, a residual concern may be that the results in Schweinberger et al., (2007) may in part have reflected a difference between the amount of information for speaker recognition available in dynamic compared to static presentations. Although facial movement has not traditionally been considered a strong cue for facial identity (Bruce & Valentine, 1988), more recent work shows that facial motion information can give participants an advantage in face recognition under certain circumstances such as visual degradation (Lander & Chuang, 2005). Moreover, experiments using more extreme visual degradation, such as point-light techniques, have shown that familiar speakers can be recognized from dynamic information only (Rosenblum, Niehus, & Smith, 2007). The effects of facial motion therefore cannot be ruled out as a major contributor to the voice recognition difference between dynamic and static visual stimuli.

The stimuli in this, and the following chapters, were presented in various different audiovisual conditions, based on the correspondence of the visual stimulus (for Experiments 1-3, participants were asked to identify the *voice*, and correspondence referred to the face presented with the target voice, while this was reversed for Experiment 4). To clarify, the target stimuli in Experiments 1-3 were presented either alone (*voice-only*) or together with faces of *corresponding* or *noncorresponding* identity. The *noncorresponding* conditions contained, firstly, faces of differing identity, but the same level of familiarity as the voice (eg. A familiar voice with a different familiar face), and these instances are referred to as the *noncorresponding-within* condition (within familiarity). The second *noncorresponding* condition contained faces of differing identity

and familiarity (eg. A familiar voice with an unfamiliar face) and these instances are referred to as the *noncorresponding-across* condition (across familiarity). The aims of Experiment 1 were to ascertain and confirm previous indications that: (1) voice recognition performance can be modulated by audiovisual condition (2) a disproportionate modulation of voice recognition performance occurs for dynamic videos in comparison to static visual presentations and that 3) voice recognition performance can be modulated by familiarity level, which may give indications that multimodal representations of person identity exist.

Method

Participants

Thirty participants (28 females, mean age 20.7 years), all in regular contact with the lecturers used as familiar speakers completed the experiment. Participants were all undergraduates of the Friedrich-Schiller University Institute of Psychology, were offered a choice of money (5€ per hour) or course credit for their participation and filled out a questionnaire indicating their level of familiarity with the familiar and unfamiliar speakers.

Stimuli and Apparatus

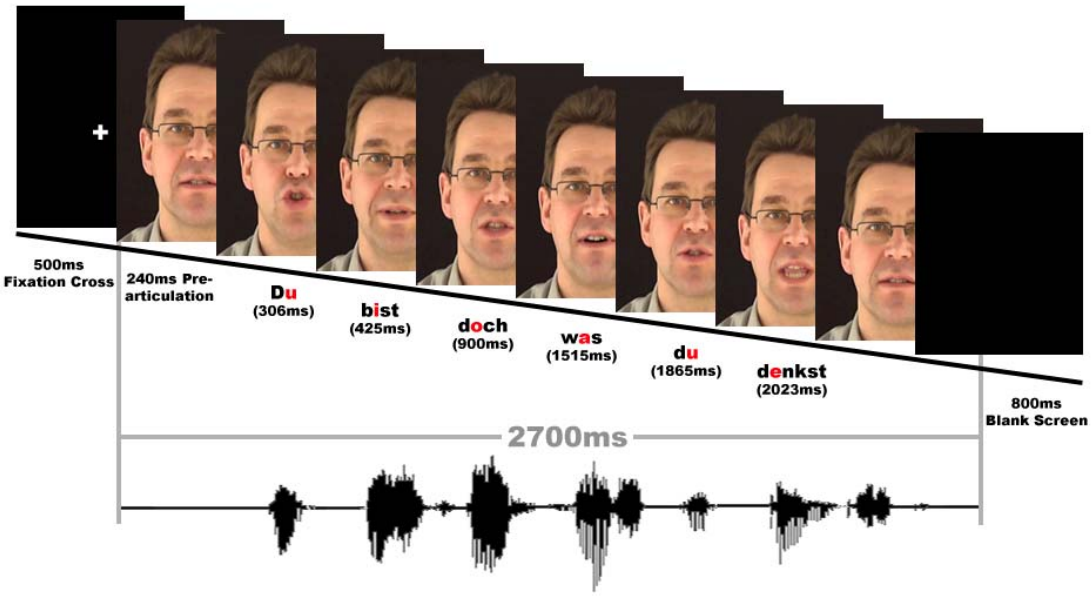


Figure 3 An example of a typical dynamic audiovisual trial, with VO timings in parenthesis. Examples of the stimuli can be viewed at http://www2.uni-jena.de/svw/Allgpsy1/stimuliexamples.html.

An example of an audiovisual trial is shown in Figure 3. The familiar speakers consisted of four Professors from the Institute for Psychology, Friedrich-Schiller University Jena – of which the participants attended at least one of their courses for at least a full semester. This amounted to approximately 13 weekly blocks of 90-minute lecturing contact with the Professors. Four other people (unfamiliar speakers) were video recorded saying the standardised sentence “Du bist doch was du denkst”. This sentence was chosen due to the number of initial stop-consonants, which made the synchronisation process simpler while still being relatively meaningful. All of the speakers were male and were matched for age.

Volunteers were trained to say the sentence with standardised timing and rhythm using a sample video clip of the target sentence. Several recordings of the sentence were subsequently taken so that there would be a wider range of choices for use at the editing stage. Video clips were recorded using a Sony DCR-DVD403E digital camcorder, and faces were evenly lit by three indirect 390W spotlights (placed out of shot, either side and above the speaker, with one central below the speaker's face), giving a luminance level of approximately 35 cd/m^2. The video clips were rendered into a 5.4 x 7 cm movie digitised at 25 frames per second (one frame = 40 ms). Static faces simply consisted of the first frame of the respective video, showing an unarticulating face for 2700 ms. Videos were presented on a computer monitor, in colour, at a viewing distance of 90 cm, which was fixed by the employment of a chin-rest. Voice clips were recorded simultaneous to the video recording, with a Sennheiser MD-421 dynamic microphone placed directly in front of the speaker, just out of camera-shot. These were digitised in Mono at 44,1 kHz with 16-bit resolution, and normalised for mean amplitude. The auditory stimuli were presented during experimentation via Sennheiser headphones. Both video and audio clips were edited to a common duration of 2700ms, considered sufficient for adequate voice recognition performance (Schweinberger, Herholz, & Sommer, 1997).

Design and Procedure

As with our past experiment, (Schweinberger et al., 2007) the instructions emphasized that participants should attentively view the visual stimuli, but should make familiar/unfamiliar responses exclusively based on the speaker's *voice*. Before the

experiment, each participant was asked to complete a questionnaire in which they verified that they were highly familiar with the familiar speakers and were completely unfamiliar with the unfamiliar speakers. In the case of the familiar speakers, they indicated how many of the familiar speaker's courses they had participated in, and an approximation of the percentage of lectures they had attended for each course (participants who indicated they had attended less than 80% of lectures for a particular familiar speaker were disqualified from testing). Seven conditions of audiovisual stimulation were presented for both familiar and unfamiliar voices, comprising 36 trials each, resulting in 7 x 2 x 36 = 504 experimental trials, which were presented in randomized order in three consecutive blocks of 168 trials each. Breaks were allowed every 84 trials. In order to acquaint the participants to the task, the experimental trials were preceded by 20 practice trials during which each experimental condition was represented at least once.

The seven audiovisual conditions were either 1) *voice only* (no visual stimulus), or voices with static faces using 2) a *corresponding* face (same speaker), 3) a *noncorresponding* face *within* the same familiarity set or 4) a *noncorresponding* face *across* familiarity sets. Conditions 5-7 were analogous to conditions 2-4 except that voices were shown with dynamic faces, eliciting a natural perception of a person speaking.

Each trial started with a fixation cross for 500 ms, followed by 2700 ms of stimulus, followed by 800 ms of blank screen. Using both index fingers, participants responded as quickly and accurately as possible in judging the familiarity of the voice presented. Half

the participants pressed the left CTRL key for familiar voices and the right CTRL key for unfamiliar voices; for the other half, this assignment was reversed. Response times (RTs) were measured relative to the onset of the auditory stimuli. Note that the identical set of auditory stimuli was used in each of the seven audiovisual conditions, such that the conditions differed only with respect to the type of additional visual stimulus a voice was combined with. Responses were scored correct if the correct key was pressed within a time window of 200-3500 ms.

Results

Reaction Time (RT) data

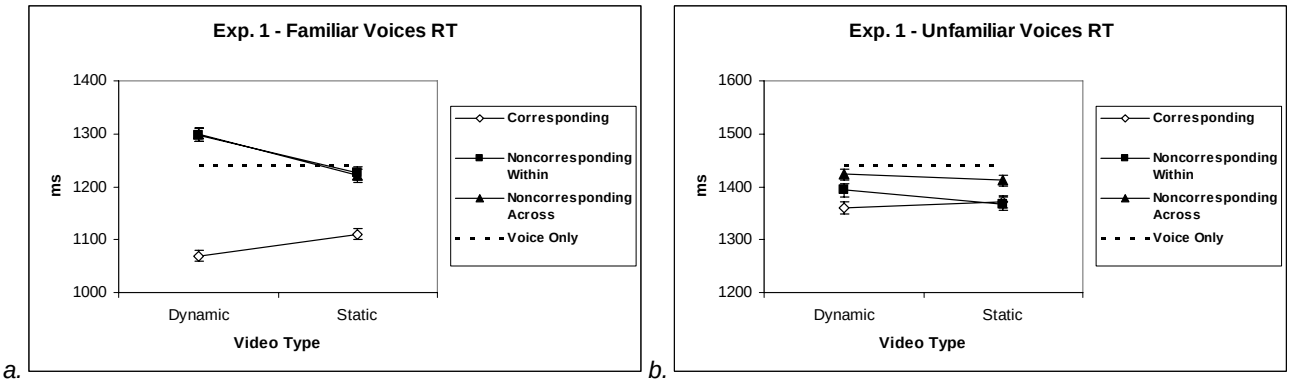


Figure 4a,b Mean response-time (RT) data for familiar and unfamiliar voices.

Figure 4*a,b* displays the mean correct RT data for Experiment 1. The data were initially submitted to an analysis of variance (ANOVA) with repeated measures for presentation condition (7 levels) and voice familiarity. Where appropriate, epsilon corrections were performed for heterogeneity of covariances (Huynh & Feldt, 1976) throughout.

In RTs, there were significant main effects of familiarity, $F(1, 29) = 69.88$, $p < 0.001$, reflecting faster responses to familiar than unfamiliar voices, and of presentation condition, $F(6, 174) = 28.80$, $p < 0.001$. Importantly, these effects were moderated by a significant interaction of familiarity by presentation condition, $F(6, 174) = 14.14$, $p < 0.001$. This interaction reflected the fact that audiovisual presentation condition had larger effects on familiar voices compared to unfamiliar voices, although presentation condition significantly affected RTs for both familiar voices, $F(6, 174) = 39.85$, $p < 0.001$, and unfamiliar voices, $F(6, 174) = 4.77$, $p < 0.001$.

In order to more clearly ascertain the voice recognition effects for audiovisual stimuli compared to voice-only stimuli, each audiovisual condition was compared with the voice-only baseline.

Familiar voices

Corresponding – Both the dynamic, $F(1, 29) = 84.66$, $p < 0.001$ and static conditions, $F(1, 29) = 56.72$, $p < 0.001$, demonstrated RT benefits in comparison with the voice-only baseline.

Noncorresponding-within – The dynamic stimuli, $F(1, 29) = 10.83$, $p < 0.001$, demonstrated significant costs compared to the baseline, but the static condition was not significantly different from voice-only, $F(1, 29) = 0.60$, $p > 0.05$.

Noncorresponding-across – The dynamic condition demonstrated significant costs relative to the baseline, $F(1, 29) = 6.85$, $p < 0.001$, but the static condition was not significantly different from voice-only performance, $F(1, 29) = 2.14$, $p > 0.05$.

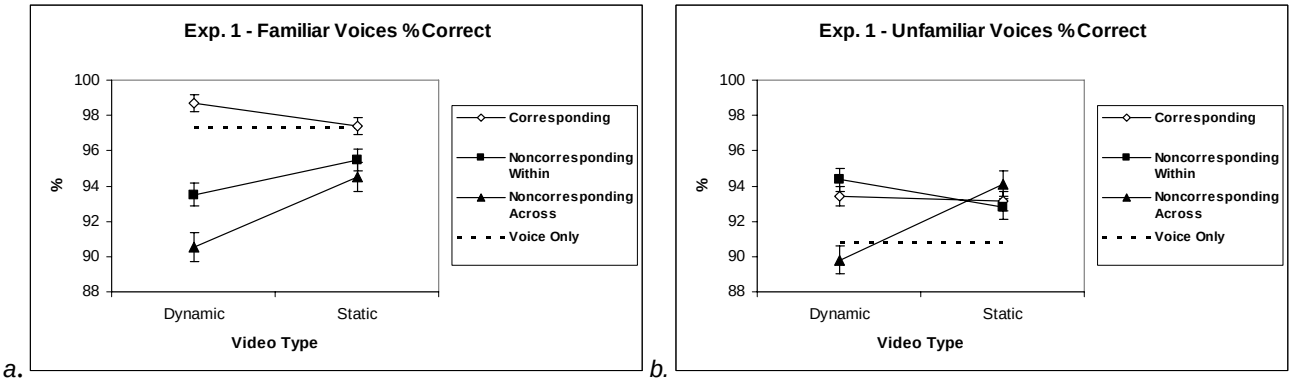
Unfamiliar voices

Corresponding – Both the dynamic, $F(1, 29) = 12.40$, $p < 0.01$, and static conditions, $F(1, 29) = 11.78$, $p < 0.05$, displayed significant benefits compared to the voice-only condition.

Noncorresponding-within – the dynamic stimuli demonstrated a nonsignificant trend for benefits, $F(1, 29) = 3.19$, $p = 0.085$, while static stimuli, $F(1, 29) = 13.55$, $p < 0.001$, showed significant benefits compared to baseline.

Noncorresponding-across – neither dynamic, $F(1, 29) = 0.70$, $p > 0.05$, or static, $F(1, 29) = 2.32$, $p > 0.05$, stimuli showed significant RT benefits compared to baseline performance.

Percentage-Correct Data



a. *b.*

Figure 5a,b Mean percentage-correct data for familiar and unfamiliar voices.

Figure 5*a,b* displays the percentage-correct data for familiar voices in Experiment 1. An analogous ANOVA was performed for response accuracies, i.e., percentages of correct responses per condition. This ANOVA yielded a trend for a main effect of familiarity, $F(1, 29) = 4.10$, $p = 0.052$, reflecting slightly more accurate responses to familiar than unfamiliar voices. There was a main effect of presentation condition, $F(6, 174) = 10.51$,

$p < 0.001$. Again, there was a significant interaction of familiarity by presentation condition $F(6, 174) = 5.16$, $p < 0.01$. The interaction again reflected that the effects of presentation condition on response accuracy were greater for familiar voices than for unfamiliar voices, but presentation condition had significant effects on familiar voices, $F(6, 174) = 11.05$, $p < 0.001$, and unfamiliar voices, $F(6, 174) = 4.23$, $p < 0.05$.

Familiar voices

Corresponding – The dynamic, $F(1, 29) = 4.39$, $p < 0.05$, and the static conditions, $F(1, 29) = 0.01$, $p < 0.01$, demonstrated significant benefits compared to baseline.

Noncorresponding-within – There were significant costs to accuracy for dynamic, $F(1, 29) = 19.59$, $p < 0.001$, and static stimuli, $F(1, 29) = 7.95$, $p < 0.01$.

Noncorresponding-across – The dynamic, $F(1, 29) = 14.02$, $p < 0.001$, and static conditions, $F(1, 29) = 5.30$, $p < 0.05$, displayed significant costs compared to baseline performance.

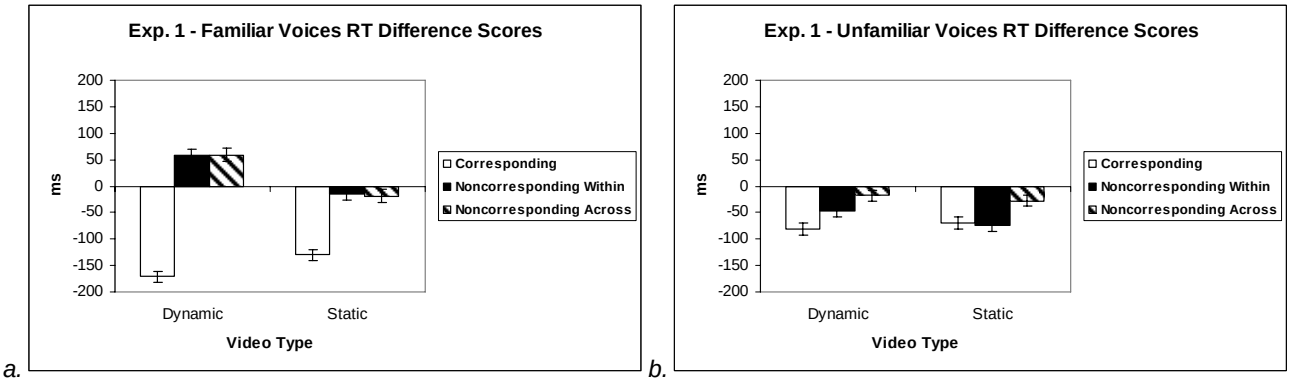
Unfamiliar voices

Corresponding – The dynamic condition was not significantly different to the voice-only condition, $F(1, 29) = 2.26$, $p > 0.05$, while the static condition displayed a nonsignificant trend for benefits compared to baseline, $F(1, 29) = 3.21$, $p = 0.084$.

Noncorresponding-within – The dynamic condition demonstrated significant benefits compared to the voice-only baseline, $F(1, 29) = 4.39$, $p < 0.05$, while the static condition was not significantly different from baseline performance, $F(1, 29) = 2.66$, $p > 0.05$.

Noncorresponding-across – The dynamic condition was not significantly different from voice-only performance, $F(1, 29) = 1.09$, $p > 0.05$, but the static condition demonstrated significant benefits to response accuracy in relation to baseline, $F(1, 29) = 9.24$, $p < 0.01$.

RT Difference Scores



a. *b.*

Figure 6a,b RT difference scores for familiar and unfamiliar voices. Difference scores were calculated by subtracting the mean for the voice-only baseline from the means of each audiovisual condition.

Because of the interaction of presentation condition and familiarity, separate analyses were performed for familiar and unfamiliar voices. In order to more systematically evaluate the effects of presentation condition, and because the benefits and costs caused by *corresponding* and *noncorresponding* faces respectively (relative to the auditory only baseline) were of primary interest, a different method of analysis was employed. These analyses were performed on the RT scores of each experimental condition minus the *voice only* baseline condition (Figure 6*a,b*). ANOVAs on these data involved repeated measures on face correspondence (*corresponding*, *noncorresponding-within*, and *noncorresponding-across*) and animation mode (*dynamic* and *static*).

Familiar Voices

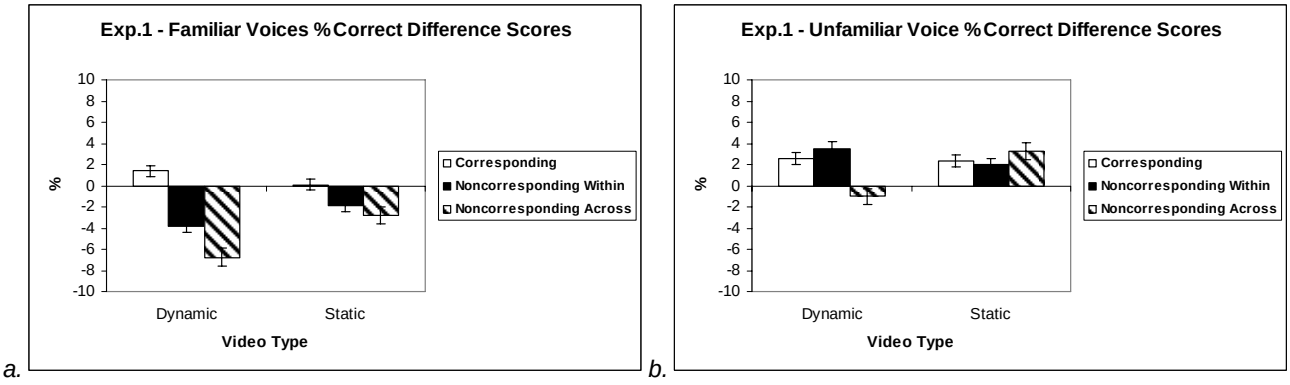
Figure 6*a* displays the RT difference scores for familiar voices in Experiment 1. The ANOVA on this data revealed main effects of correspondence, $F(2, 58) = 65.33$, $p < 0.001$, and animation, $F(1, 29) = 10.53$, $p < 0.01$. There was also a significant interaction found for correspondence by animation, $F(2, 58) = 17.85$, $p < 0.001$. This interaction indicated highly significant RT benefits for the *corresponding* condition compared to the *noncorresponding-within*, $F(1, 29) = 73.06$, $p < 0.001$, and *noncorresponding-across*, $F(1, 29) = 74.95$, $p < 0.001$, conditions. The two *noncorresponding* conditions were not significantly different from each other, $F(1, 29) = 0.04$, $p > 0.05$. In the *corresponding* condition, larger RT benefits were observed for dynamic as compared to static face presentations, $F(1, 29) = 6.73$, $p < 0.05$. In the *noncorresponding-within* condition, larger RT costs were seen for dynamic as compared to static face presentations, $F(1, 29) = 17.80$, $p < 0.001$. The *noncorresponding-across* condition produced large RT costs for dynamic compared to static face presentations, $F(1, 29) = 18.06$, $p < 0.001$. In fact, as can be seen in Figure 6a, the costs to RT performance for static stimuli were non-existent.

Unfamiliar Voices

Figure 6*b* displays the RT difference scores for unfamiliar voices in Experiment 1. Analogous analyses for unfamiliar voices revealed a main effect of correspondence, $F(2, 58) = 5.26$, $p < 0.01$, but animation was not significant, $F(1, 29) = 0.52$, $p > 0.05$, and there was no interaction, $F(2, 34) = 2.10$, $p > 0.05$. The *corresponding* condition displayed

significantly greater RT benefits compared to the *noncorresponding-across* condition, $F(1, 29) = 11.19$, $p < 0.01$, but not in comparison with the *noncorresponding-within* condition, $F(1, 29) = 0.80$, $p > 0.05$. The *noncorresponding-within* condition displayed significant benefits compared to the *noncorresponding-across* condition, $F(1, 29) = 4.62$, $p < 0.05$. Importantly, and in striking contrast to the results for familiar voices, the dynamic versus static comparisons within each correspondence condition were not significant, $Fs(1, 29) < 2.2$, $ps > 0.1$.

Percentage-Correct Difference Scores



a. *b.*

Figure 7a,b Percentage-Correct difference scores for familiar and unfamiliar voices.

Familiar Voices

Figure 7*a* displays the percentage-correct difference scores for familiar voices in Experiment 1. Analogous ANOVAs were performed on the percent correct difference scores of each experimental condition minus the *voice only* baseline condition, using the same factors as for the RT difference scores above. The ANOVA for familiar voices revealed main effects of correspondence, $F(2, 58) = 12.53$, $p < 0.001$, and animation, $F(1, 29) = 9.96$, $p < 0.01$. A significant interaction between these two factors was also found, $F(2, 58) = 8.14$, $p < 0.01$. This was characterised by benefits in accuracy for the *corresponding* condition relative to both the *noncorresponding-within* condition, $F(1, 29) = 23.94$, $p < 0.001$, and the *noncorresponding-across* condition, $F(1, 29) = 18.99$, $p < 0.001$. The two *noncorresponding* conditions were not significantly different from each other, $F(1, 29) = 2.35$, $p > 0.05$. There were benefits to response accuracy for dynamic compared to static presentations in the *corresponding* condition, $F(1, 29) = 6.90$, $p < 0.01$. The *noncorresponding-within* condition displayed costs for dynamic compared to static face presentations, $F(1, 29) = 5.59$, $p < 0.05$, and the *noncorresponding-across* condition

also showed large costs in accuracy for dynamic compared to static face presentations, $F(1, 29) = 10.31$, $p < 0.01$.

Unfamiliar Voices

Figure 7*b* displays the percentage-correct difference scores for unfamiliar voices in Experiment 1. The ANOVA on this data demonstrated no main effect of correspondence, $F(2, 58) = 1.86$, $p > 0.05$ and a nonsignificant trend for a main effect of animation, $F(1, 29) = 3.49$, $p = 0.072$. There was however a significant interaction for the two conditions, $F(2, 58) = 9.32$, $p < 0.01$. Comparing the correspondence conditions to each other yielded no significant comparisons, Fs$(1, 29) < 3$, $p > 0.05$. For the dynamic versus static comparisons, *noncorresponding-across* was the only condition to show a significant difference between the conditions of animation, $F(1, 29) = 14.67$, $p < 0.001$, where there were costs to voice recognition performance for the dynamic stimuli in comparison to the static benefits. The *noncorresponding-within* condition demonstrated a nonsignificant trend for benefits of dynamic compared to static stimuli, $F(1, 29) = 3.13$, $p = 0.088$.

Discussion

Firstly, it is notable that effects of AVI appear to be prevalent in this experiment. Compared directly to the voice-only baseline, the majority of the audiovisual conditions result in significantly altered performance. This is particularly clear for the dynamic stimuli where, in general, familiar *corresponding* stimuli result in significantly better performance than in the voice-only condition, while dynamic *noncorresponding* stimuli generally result in slower and less accurate performance compared to the voice-only baseline. Facilitation effects appear to be strongest when dynamic familiar stimuli are involved, although familiar static *corresponding* stimuli also result in small facilitations of performance compared to the voice-only baseline. It may be the case that the small facilitation in performance brought about by static face presentations reflects strategic cue usage, that is, since the face is recognised more quickly than the voice, it is possible that the face creates strategic expectancies for the identity of the voice, as alluded to in Schweinberger et al., (2007). This explanation seems more likely due to the larger facilitation effects afforded by familiar dynamic *corresponding* stimuli, which may be indicative of AVI. Furthermore, responses to static *noncorresponding* stimuli, particularly for familiar voices, and unfamiliar voices with a familiar face, were generally similar to voice-only performance. Taken together with the data on dynamic *noncorresponding* performance, it appears that static *noncorresponding* stimuli can be ignored, while dynamic *noncorresponding* stimuli cannot, resulting in poorer voice recognition performance for the latter. These differences between the facilitation and inhibition effects of dynamic and static stimuli in voice recognition, strongly suggest that AVI is an important factor in person perception.

Experiment 1 revealed voice recognition benefits for *corresponding* familiar stimuli in comparison to other stimuli. Importantly, dynamic-*corresponding* stimuli resulted in significantly large benefits in RTs compared to static stimuli. Response accuracy was generally high across all of the conditions, especially for familiar voices with *corresponding* faces. It is thus possible that a ceiling effect accounts for the absence of significant differences between dynamic and static stimuli in the *corresponding* condition.

A converse pattern was apparent for the *noncorresponding* conditions. That is, in the conditions where voices were presented with a face of different identity, voice recognition was slower and less accurate. This was particularly the case for familiar voices, where the *noncorresponding-within* (voices with a different familiar face) and the *noncorresponding-across* (voices with an unfamiliar face) conditions showed similar degradations in response time for dynamic stimuli presentations. For familiar voices, dynamic *noncorresponding* faces had the effect of slowing response times compared to static *noncorresponding* faces, where response times were almost identical to the voice-only baseline. Percentage-correct data for the *noncorresponding-within* condition displayed larger costs to voice recognition accuracy in comparison to the *corresponding* condition, and the *noncorresponding-across* condition displayed even larger costs. For both *noncorresponding* conditions, these costs were much more pronounced for dynamic compared to static presentations, which suggests that AVI is interfering with performance in this case. These results are largely in line with previous findings (Schweinberger et al. 2007).

Perhaps most notable are the differences in familiar voice recognition performance between dynamic and static stimuli. When a static face with the same identity as the voice is presented, it leads to faster and more accurate performance in comparison to the voice-only baseline. When the static face is of a different identity to the voice, the reaction times are similar to baseline, while costs to accuracy are incurred. These benefits and costs perhaps reflect a strategic cue for voice recognition (Rosenblum et al., 1996), although this seems unlikely given the negligible costs between the *noncorresponding* conditions. For instance, if a static familiar face cued recognition of a voice as familiar, responses when a familiar face of different identity (as in the *noncorresponding-within* condition) is presented, should be more similar to the *corresponding* condition than the *noncorresponding-across* condition.

The *corresponding* dynamic stimuli may lead to the fastest and most accurate responses because they are more realistic than the other stimuli. The *noncorresponding* conditions likely result in greater costs for the dynamic stimuli compared to the other stimuli because the “cognitive compellingness” (Warren et al., 1981) of the audiovisual pairing means that the face cannot be ignored in making the voice recognition response. The illusion that the voice is being spoken by the moving face causes the stimuli to be perceived as belonging to the same event, causing them to be integrated. The initial perception of the two stimuli as belonging to the same event, may cause confusion in making the response decision, which gives rise to delays and errors. The data for unfamiliar voices follows a less discernable pattern. What is most notable, however, is that the *noncorresponding-across* condition incurs small costs to accuracy when dynamic stimuli are presented, compared to the apparent benefits of any visual stimuli presentation for the other conditions. It

should also be noted that that it seems that the pairing of an unfamiliar voice with any unfamiliar face, causes an improvement in performance. This suggests that the physical constraints afforded by voices in allowing unfamiliar face-voice matching (Kamachi et al., 2003) are not necessarily prevalent in these circumstances. It seems more likely that where there is perceptual uncertainty about the familiarity of a voice, the presence of any unfamiliar face may act as a strong cue to respond “unfamiliar”. Importantly, the presence of a dynamic familiar face appears to have a markedly detrimental effect on the correct identification of an unfamiliar voice. This might suggest that unfamiliar voices are perceived as more familiar when they are presented in synchrony with a familiar face. However, it is also conceivable that the presence of a dynamic familiar face interferes with the participant’s decision, making it more difficult to correctly respond “unfamiliar”.

Overall, these effects are reminiscent of the McGurk effect (McGurk & MacDonald, 1976), in that discrepant, synchronised visual stimuli apparently cannot be ignored. Rather, synchronised faces have a stronger influence on the perception of auditory stimuli, despite the participants having been expressly instructed to base their responses on what they heard. A possible reason for the difference between the dynamic and static conditions may be that the moving visual stimuli result in a perception that is more closely related to real-life. Lander and colleagues have found that moving faces are more quickly and accurately recognised than static faces (Lander & Bruce, 2000; Lander et al., 2005), so it is conceivable that the dynamic and static effects occur, due to the extra identity information available to the participants. Benefits to face recognition for moving faces compared to static faces have normally been shown with degraded stimuli (Rosenblum et al., 2006; Lachs et al., 2004a; Lachs et al., 2004b). Although the stimuli used in the

current experiment were clearly seen and heard, the possibility that dynamic videos act as more salient and information-rich cues to the identity of the voice, cannot be discounted. Therefore, it may not be completely clear whether time-synchronisation underlie the AVI effects observed in Experiment 1, or whether facial motion *per se* is the primary factor. This possibility will be investigated in Experiment 2.

EXPERIMENT 2

Introduction

Experiment 2 used the same stimuli as Experiment 1, with the exception that the static condition was replaced by a backwards-video condition. Moving faces have previously been shown be more quickly and efficiently recognised than static pictures (Lander & Chuang, 2005). Therefore it is necessary to compare two dynamic face conditions, as the differences shown in Experiment 1 may be largely due to the extra identity information afforded by the dynamic videos in comparison to the information available in the static pictures. It is conceivable that the extra facial information conveyed by moving faces provides a stronger strategic cue than a static face, and it is this which causes the patterns of costs and benefits previously suggested as evidence for AVI in voice recognition. In an attempt to disprove this possible hypothesis, backwards videos were used as a second condition of facial motion, where the motion information contained in the videos were the same as the forwards videos, only the frame order had been reversed. If facial motion *per se* accounts for the differences between dynamic and static stimuli seen in Experiment 1, there should be no differences between forwards and backwards conditions. By contrast, if the effects for forwards and backwards conditions are identical, it might indeed be the case that the information-content in moving stimuli result in the observed effects, that is, that the effects seen in Experiment 1 were primarily due to the extra information available in dynamic presentations of faces in comparison to static face presentations.

Method

Participants

Twenty participants (20 females, mean age 20.2 years), all in regular contact with the lecturers used as familiar speakers completed the experiment. Participants were all undergraduates of the Friedrich-Schiller University Institute of Psychology, were offered a choice of money (5€ per hour) or course credit for their participation, and filled out a questionnaire indicating their level of familiarity with the familiar and unfamiliar speakers.

Stimuli and Apparatus

Experiment 2 used the same stimuli and apparatus as Experiment 1 with the exception that the static stimuli were omitted and backwards-videos were added. Backwards videos were made simply by reversing the frame order of the dynamic (synchronised) videos used in Experiment 1.

Design and Procedure

The experiment used the same instructions as Experiment 1, that participants should attentively view the visual stimuli, but should make familiar/unfamiliar responses exclusively based on the speaker's *voice*. Seven conditions of audiovisual stimulation were presented for both familiar and unfamiliar voices, comprising 36 trials each, resulting in 7 x 2 x 36 = 504 experimental trials, which were presented in randomized order in three consecutive blocks of 168 trials each. Breaks were allowed every 84 trials. In order to

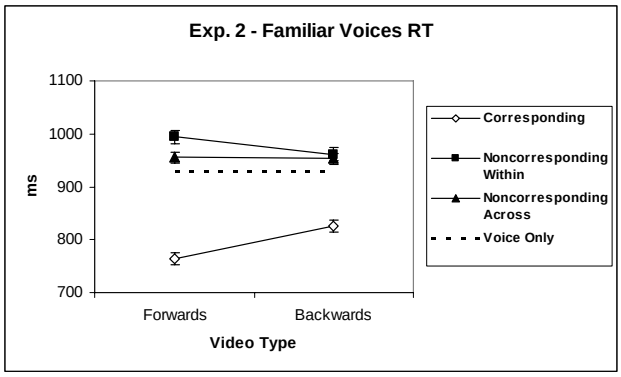
acquaint the participants to the task, the experimental trials were preceded by 20 practice trials during which each experimental condition was represented at least once.

The seven audiovisual conditions were either 1) *voice only* (no visual stimulus), or voices with *backwards* videos using 2) a *corresponding* face (same speaker), 3) a *noncorresponding* face *within* the same familiarity set (i.e., for a familiar voice, a different familiar face was shown, and for an unfamiliar voice, a different unfamiliar face was shown), or 4) a *noncorresponding* face *across* familiarity sets (e.g., for a familiar voice, an unfamiliar face was shown). Conditions 5-7 were analogous to conditions 2-4 except that voices were shown with forwards-playing videos, eliciting a natural perception of a person speaking. The temporal attributes of the trials remained the same as in Experiment 1 and responses were measured in the same way.

Results

Reaction Time (RT) data

a.



b.

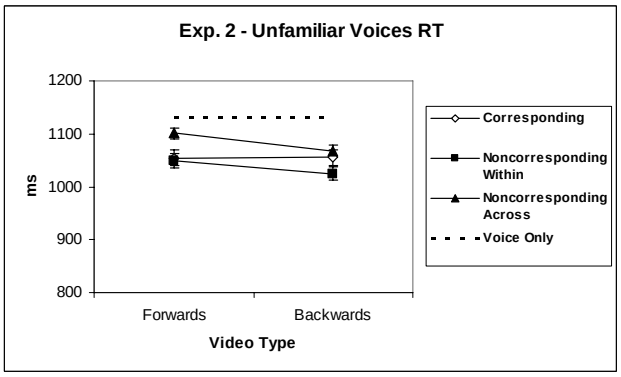


Figure 8a,b Mean reaction-time (RT) data for familiar and unfamiliar voices.

Figure 8*a,b* display the mean correct RT data for Experiment 2. This data was initially submitted to an analysis of variance (ANOVA) with repeated measures for presentation condition (7 levels) and voice familiarity. Where appropriate, epsilon corrections were performed for heterogeneity of covariances (Huynh et al., 1976) throughout.

In RTs, we found a significant main effect of familiarity, $F(1, 19) = 38.21$, $p < 0.001$, reflecting faster responses to familiar than unfamiliar voices, and of presentation condition, $F(6, 114) = 18.45$, $p < 0.001$. These effects were moderated by a significant interaction of familiarity by presentation condition, $F(6, 114) = 22.77$, $p < 0.001$. This interaction reflected the fact that the effects of audiovisual presentation condition were significantly larger for familiar voices than for familiar voices, although presentation condition significantly affected RTs for familiar voices, $F(6, 114) = 39.27$, $p < 0.001$, and unfamiliar voices, $F(6, 114) = 5.57$, $p < 0.001$.

In order to more clearly ascertain the voice recognition effects for audiovisual stimuli compared to voice-only stimuli, each audiovisual condition was compared with the voice-only baseline.

Familiar Voices

Corresponding – Forwards, $F(1, 19) = 93.38$, $p < 0.001$, and backwards, $F(1, 19) = 28.12$, $p < 0.001$, videos displayed significant RT benefits in comparison with the voice-only baseline.

Noncorresponding-within – In this case, the forwards, $F(1, 19) = 15.49$, $p < 0.001$, and backwards, $F(1, 19) = 5.23$, $p < 0.001$, videos demonstrated significant costs compared to the baseline.

Noncorresponding-across – Forwards and backwards stimuli did not display significant differences from baseline performance, $F(1, 19) = 2.55$, $p > 0.05$, and $F(1, 19) = 1.71$, $p > 0.05$, respectively.

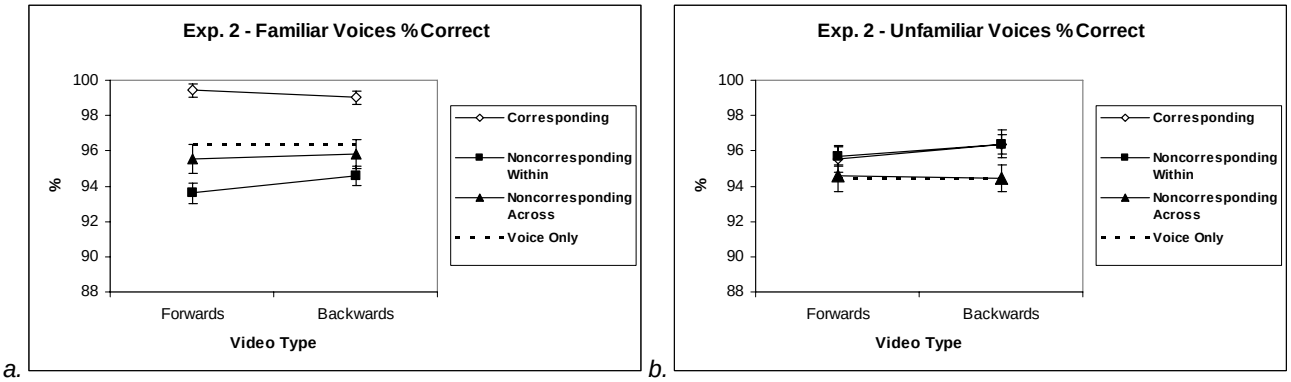
Unfamiliar Voices

Corresponding – Forwards, $F(1, 19) = 14.47$, $p < 0.01$, and backwards, $F(1, 19) = 7.55$, $p < 0.05$ videos resulted in benefits compared to the voice-only condition.

Noncorresponding-within – The forwards, $F(1, 19) = 20.45$, $p < 0.001$, and backwards, $F(1, 19) = 21.62$, $p < 0.001$, videos also showed significant benefits compared to the voice-only baseline.

Noncorresponding-across - The forwards condition was not significantly different from baseline, $F(1, 19) = 1.50$, $p > 0.05$, but the backwards, $F(1, 19) = 6.57$, $p < 0.001$, condition showed significant RT benefits compared to baseline performance.

Percentage-Correct data



a. *b.*

Figure 9a,b Mean percentage-correct data for familiar and unfamiliar voices.

Figure 9*a,b* display the percentage-correct data for Experiment 2. An analogous ANOVA was performed for the percentage-correct data, the accuracy of responses in each condition. Presentation condition was found to be significant, $F(6, 114) = 5.12$, $p < 0.001$ but familiarity was not, $F(1, 19) = 1.14$, $p > 0.05$. Again, there was a significant interaction for familiarity by presentation condition, $F(6, 114) = 3.94$, $p < 0.01$. The effect of presentation condition on response accuracy was significant for familiar voices, $F(6, 114) = 7.08$, $p < 0.001$, but not for unfamiliar voices, $F(6, 114) = 1.39$, $p > 0.05$.

Familiar voices

Corresponding – Forwards, $F(1, 19) = 25.83$, $p < 0.01$, and backwards, $F(1, 19) = 12.73$, $p < 0.01$, videos demonstrated significant response accuracy benefits compared to baseline.

Noncorresponding-within – The forwards, $F(1, 19) = 4.22$, $p = 0.054$, and backwards, $F(1, 19) = 3.18$, $p = 0.091$, videos demonstrated a nonsignificant trend for costs to response accuracy in comparison to baseline.

Noncorresponding-across – The forwards and backwards conditions were not significantly different to baseline performance, $Fs(1, 19) < 1$, $ps > 0.05$.

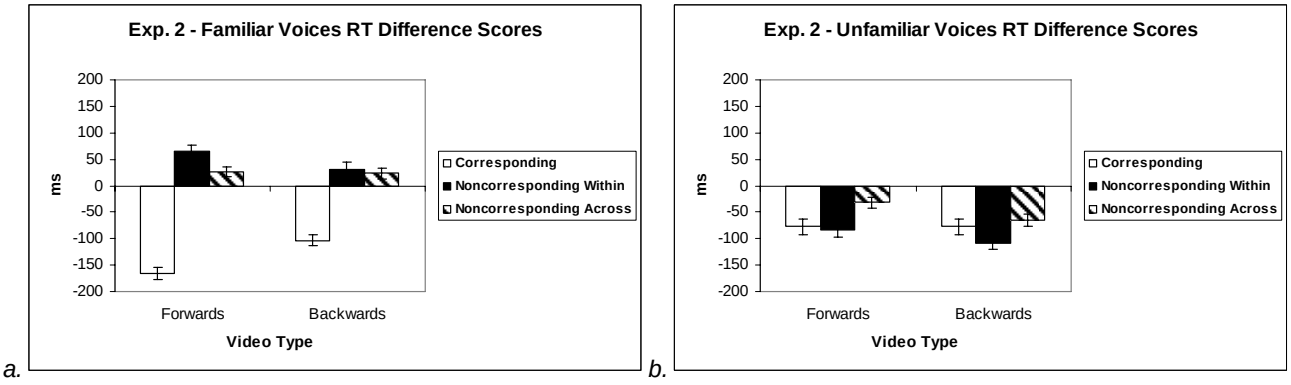
Unfamiliar voices

Corresponding – The forwards condition was not significantly different to the voice-only baseline, $F(1, 19) = 0.88$, $p > 0.05$, while the backwards condition displayed a nonsignificant trend for benefits compared to baseline performance, $F(1, 19) = 4.03$, $p = 0.059$.

Noncorresponding-within – Once again, the forwards condition was not significantly different from the voice-only baseline, $F(1, 19) = 1.26$, $p > 0.05$, but the backwards condition demonstrated significant benefits, $F(1, 19) = 5.14$, $p < 0.05$.

Noncorresponding-across – The forwards and backwards conditions were not significantly different to the voice-only baseline, $Fs(1, 19) < 1$, $ps > 0.05$.

RT difference scores



a. *b.*

Figure 10a,b RT difference scores for familiar and unfamiliar voices.

As in Experiment 1, RT difference scores were calculated by subtracting the *voice only* baseline condition from each experimental condition (Figure 10*a,b*). ANOVAs on these data involved repeated measures on animation mode (*forwards* vs. *backwards*), and face correspondence (*corresponding*, *noncorresponding-within*, and *noncorresponding-across*).

Familiar Voices

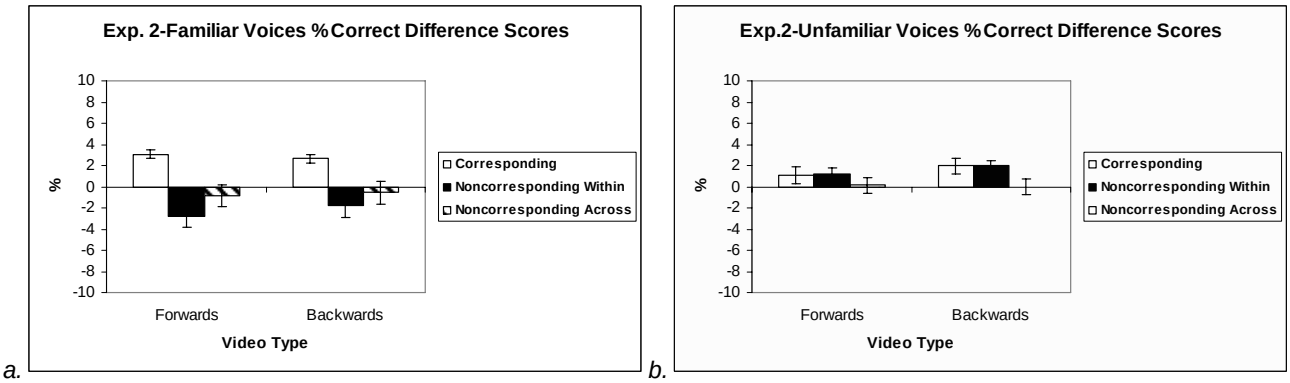
Figure 10*a* displays the RT data for familiar voices in Experiment 2. The ANOVA on this data revealed main effects of correspondence, $F(2, 38) = 67.41$, $p < 0.001$, but not animation, $F(1, 19) = 1.36$, $p > 0.05$. There was a significant interaction between the two factors, $F(2, 38) = 7.94$, $p < 0.001$. *Corresponding* stimuli resulted in significantly large benefits compared to the costs observed for *noncorresponding-within*, $F(1, 19) = 79.50$, $p < 0.001$, and *noncorresponding-across*, $F(1, 19) = 75.88$, $p < 0.001$, conditions. The

noncorresponding-within condition displayed a nonsignificant trend for larger RT costs compared to the *noncorresponding-across* condition, $F(1, 19) = 4.33$, $p = 0.051$. Larger RT benefits were demonstrated for forwards videos than for backwards videos, $F(1, 19) = 15.75$, $p < 0.001$. In the *noncorresponding-within* condition RT costs tended to be larger for the forwards stimuli in comparison to backwards stimuli, but this comparison did not quite reach significance, $F(1, 19) = 3.64$, $p = 0.071$. Comparing forwards and backwards conditions in the *noncorresponding-across* condition, $F(1, 19) = 0.03$, $p > 0.05$, yielded no significance.

Unfamiliar Voices

Figure 10*b* displays the RT difference scores for unfamiliar voices in Experiment 2. The ANOVA revealed that there was a main effect of correspondence, $F(2, 38) = 4.07$, $p < 0.05$, but not for animation $F(1, 19) = 2.44$, $p > 0.05$ and there was no significant interaction for the two factors, $F(2, 38) = 1.04$, $p > 0.05$. *Noncorresponding-within* versus *noncorresponding-across* was the only significant comparison between correspondence conditions, $F(1, 19) = 9.19$, $p < 0.01$, where *noncorresponding-within* resulted in significantly larger costs. The *corresponding* condition displayed a nonsignificant trend for larger benefits than the *noncorresponding-across* condition, $F(1, 19) = 9.19$, $p = 0.089$. Of the three conditions of correspondence, only *noncorresponding-across* displayed a significant effect of animation, $F(1, 19) = 4.51$, $p < 0.05$, where larger costs were incurred by forwards compared to backwards videos.

Percentage-Correct Difference Scores



a. *b.*

Figure 11a,b Mean percentage-correct difference scores for familiar and unfamiliar voices.

Familiar Voices

Figure 11*a* displays the percentage-correct difference scores for familiar voices in Experiment 2. Analogous ANOVAs were performed on the percent-correct difference scores as previously. The ANOVA for familiar voices revealed a main effect of correspondence, $F(2, 38) = 19.86$, $p < 0.001$, but not animation, $F(1, 19) = 0.14$, $p > 0.05$ and there was no significant interaction, $F(2, 38) = 0.32$, $p > 0.05$. The *corresponding* condition resulted in large benefits to performance accuracy compared to the costs observed for *noncorresponding-within*, $F(1, 19) = 32.88$, $p < 0.001$, and *noncorresponding-across*, $F(1, 19) = 27.17$, $p < 0.001$, conditions. The *noncorresponding-within* condition displayed a trend for larger costs compared to the *noncorresponding-within* condition, $F(1, 19) = 3.09$, $p = 0.095$. None of the comparisons of forwards versus backwards video presentations were significant for any of the three correspondence conditions.

Unfamiliar Voices

Figure 11*b* displays the percentage-correct difference scores for unfamiliar voices in Experiment 2. The ANOVA for unfamiliar voices shows no significant main effects of correspondence, $F(2, 38) = 2.22$, $p > 0.05$, or animation, $F(1, 19) = 1.34$, $p > 0.05$. None of the other comparisons were significant.

Discussion

The facilitation and inhibition of performance for audiovisual conditions in comparison to the voice-only baseline may further indicate that AVI is a major factor in person perception. Similar to the data previously shown for dynamic stimuli, familiar forward *corresponding* stimuli resulted in a significant facilitation of performance, while familiar forwards *noncorresponding-within* stimuli generally resulted in an inhibition of performance. Backwards *corresponding* stimuli also resulted in facilitation effects, while the backwards *noncorresponding-within* condition showed significant inhibition of performance. The observation that the effects for familiar forwards stimuli are stronger than for the backwards stimuli may be indicative of AVI, that is, that forwards *noncorresponding* stimuli are more difficult to ignore than backwards *noncorresponding* stimuli. It is important to note however, that *noncorresponding-across* stimuli did not demonstrate clear effects of AVI as in Experiment 1. In general, for both forwards and backwards presentation, performance for *noncorresponding-across* stimuli was similar to the voice-only baseline. This unexpected result is difficult to interpret but may be conceivable be due to the nature of the stimuli presented. It may be the case that the effects of Experiment 1 were so strong because a mixture of static and dynamic stimuli were presented. Considering that the effects may be partly due to the greater information content in moving stimuli (Lander et al., 2005), the effects of Experiment 1 may have caused exaggerated facilitation and inhibition of performance due to the special status of dynamic stimuli in that context. This seems unlikely however, since the other conditions of correspondence in Experiment 2 show qualitative similarities to, though faster and more accurate performance than, Experiment 1. It seems to be the case, in Experiment 2, that

when an unfamiliar face is presented with a familiar voice, it is easier to ignore. Further research, perhaps with a larger sample of participants and stimuli, is required to investigate whether this is a general theme with regards to dynamic presentations of stimuli in this context.

In Experiment 2, responses to familiar voices showed some similarities to Experiment 1. Response accuracy was high and a ceiling effect caused by the ease of the task may again account for the lack of significant differences between forwards and backwards stimuli presentations. As before, differences were seen between the three conditions of correspondence, with the *corresponding* condition bringing faster and more accurate performance and the *noncorresponding* conditions displaying slower and less accurate responses. The *corresponding* forwards videos displayed RT benefits in comparison to backwards videos. The RT effects of *noncorresponding-within* stimuli were similar to Experiment 1, in that responses were slower for forwards stimuli in comparision to backwards stimuli, though this effect represented only a nonsignificant trend. Response time and accuracy were near identical for forwards and backwards presentations in the *noncorresponding-across* condition and were similar to baseline. It seems that in this case, unfamiliar faces did not affect the recognition of familiar voices. This finding is difficult to interpret, but may reflect the high degree of familiarity with the familiar voices.

In the cases of the familiar *corresponding* and *noncorresponding-within* conditions, the forwards versus backwards effects were similar to the dynamic versus static effects displayed in Experiment 1. Most pertinently perhaps, the *noncorresponding-within* condition suggests that *noncorresponding* backwards videos are treated similarly to static

faces, that they can perhaps be ignored, while forwards videos cannot. The RT data for unfamiliar voices showed that the only significant forwards versus backwards comparison was in the *noncorresponding-across* condition, that is, where an unfamiliar voice was presented with a familiar face. Responses were faster for backwards compared to forwards faces, suggesting that forwards videos interfere more with voice recognition than the backwards presentations. Participants often described perceiving a familiar voice synchronised with a different familiar face (presented in the *noncorresponding-within* condition) as particularly striking (or humourous), and this is particularly reflected in the RT data.

Although the backwards videos often result in costs and benefits in the same direction as the forwards videos, the effects of backwards videos are often considerably smaller. It may be the case, particularly for familiar stimuli in the *noncorresponding-within* condition, that backwards videos do not form as persuasive a unitary percept as for the forwards videos. This again raises the question of whether it is the cognitive compellingness (Warren et al., 1981) of the forwards (and therefore synchronised) stimuli compared to the backwards stimuli that cause the benefits and costs displayed in this experiment. The synchrony of the speech in the forwards condition creates the perception that the moving face is the source of the heard voice. In the backwards condition, it may be the case, as for the static condition in Experiment 1, that the two modalities can be perceived more easily as separate stimuli, and as a result, do not cause such pronounced benefits and costs to voice recognition performance. It is therefore unlikely that the patterns of benefits and costs reflect simply that the identity information is richer for moving faces compared to static pictures as face recognition with dynamic stimuli might suggest (Lander et al.,

2000; Lander et al., 2005). It is more likely that the “cognitive compellingness” of the synchronised stimuli causes a stronger audiovisual percept, facilitating performance when the modalities are of matching identity, but perhaps due to this efficiency of this process, inhibits performance when the modalities are discrepant with respect to identity. These results are in line with previous research on the importance of the synchronicity of modalities in creating an audiovisual percept (Welch & Warren, 1980; Koppen & Spence, 2007; van Atteveldt, Formisano, Blomert, & Goebel, 2007; Doesburg, Emberson, Rahi, Cameron, & Ward, 2008).

Some of the differences between forwards and backwards conditions are smaller in Experiment 2 than the differences seen between the dynamic and static conditions in Experiment 1. A certain level of synchrony is seen as a fundamental requirement of AVI, but there is an expanding area of research on the importance of synchrony and a time-window of asynchrony tolerance in audiovisual phenomena (Koppen et al., 2007; van Atteveldt et al., 2007; Doesburg et al., 2008; van Wassenhove et al., 2007). A possible reason for the various nonsignificant differences between many of the forwards versus backwards comparisons may be that the forwards and backwards video may not be optimally comparable. The backwards stimuli, while representing facial movements, as it was created by reversing the same sentence as the forwards stimuli, provides only arbitrary asynchrony between the modalities. It may be the case that certain visual utterances in the backwards stimuli occur temporally close to utterances in the auditory stimulus so that it may not exert effects that are perceptually different from the forwards videos. It is relatively well-established that audiovisual speech perception effects are still attainable with substantial asynchronies (Munhall et al., 1996; van Wassenhove et al.,

2007). A more rigid and mechanical manipulation of the synchrony of the modalities would help to investigate the importance of synchrony to the effects so far shown for dynamic stimuli, and perhaps suggest a temporal window of integration, as has previously been shown for the McGurk effect (van Wassenhove et al., 2007).

CHAPTER 3:

ASYNCHRONY TOLERANCE FOR AUDIOVISUAL INTEGRATION DURING VOICE RECOGNITION

EXPERIMENT 3

Introduction

Synchronisation has been shown to be an important factor in showing audiovisual effects, particularly for simple stimuli, such as those used during spatial localisation of a sound (Soto-Faraco, Lyons, Gazzaniga, Spence, & Kingstone, 2002). For more complex stimuli, such as speech, absolute synchrony of the modalities is not considered to be as crucial to achieving effects associated with AVI. It is thought to be the case that information-rich stimuli are more readily integrated so that, to a certain extent, asynchrony is not always a significant hindrance (Calvert et al., 1998).

Originally, Munhall et al., (1996), and more recently van Wassenhove et al., (2007) tested the temporal tolerance of the McGurk effect under different conditions of asynchrony. Munhall et al. (1996) tested the McGurk illusion under audiovisual asynchronies from 360ms auditory-lead to 360ms auditory-lag in 60 millisecond steps. They found that although the participants recorded less McGurk responses when the auditory stimuli lead the visual stimuli, there was tolerance for asynchrony between approximately 60ms auditory-lead and 240ms auditory-lag. Using smaller step-sizes (33ms), van Wassenhove et al. (2007) conducted a similar study into the asynchrony tolerance of the McGurk illusion. They found that McGurk responses were prevalent in a 200ms time-window (30 milliseconds auditory lead and 170 milliseconds auditory lag (-30 - +170ms)). The

temporal window of integration has been suggested by some to be flexible (Navarra et al., 2005), in that it can be widened and shortened, while (van Wassenhove et al., 2007) suggest that it is more likely that it remains constant and can be shifted to a certain extent. The different theories have a similar basis, that is, that perceived audiovisual synchrony can be adaptable. By repeated presentation of asynchronous stimuli, perception of synchrony and asynchrony can be altered so that, for example, just-noticeable-differences between the modalities can be perceived at stimulus onset asynchronies that are either closer or farther from synchrony (Vatakis, Navarra, Soto-Faraco, & Spence, 2008). Due to the observations that the McGurk effect occurs within a defined temporal window of asynchrony, a demonstration of a similar window for voice recognition may support the suggestion that AVI is the major factor in the previous benefits and costs observed for synchronised dynamic face presentations. That is, if the unity assumption is important for the findings, then the effects should weaken as the stimuli progressively move further from synchrony. If the effects seen previously are accounted for by the relative identity-cue strengths of dynamic versus static stimuli, then it should be expected that there will be no great differences for asynchronies outside of the possible temporal window compared to synchronous stimuli.

Method

Participants

Twenty-four participants (21 females, mean age 21.4 years), who were all in regular contact with the lecturers used as familiar speakers, as outlined in Experiment 1,

completed the experiment. Participants were all undergraduates of the Friedrich-Schiller University Institute of Psychology, were offered a choice of money (7€ for 1 hour and twenty minutes) or course credit for their participation and filled out a questionnaire indicating their level of familiarity with the familiar and unfamiliar speakers.

Stimuli and Apparatus

Experiment 3 used similar stimulus preparation as the previous experiments. The visual and auditory stimuli were altered to make the synchrony adjustable. The visual stimuli were initially edited to remove the 240ms pre-articulation movements so that the first frame was the onset of the sentence. Sixteen static, unarticulating frames (640ms) were added before and after speech movements. The visual stimuli therefore, were the same across all conditions. The auditory stimuli, for the purposes of the synchronous stimuli, were edited in a similar fashion. The initial 240ms silence was deleted so that the clip began immediately with the utterance. Thereafter, 640ms windows of silence were added at the beginning and end of the auditory utterance. The new clips were therefore 640+2460+640 = 3740ms in length. To create asynchrony, clips were created by altering the onset of the auditory clips only, moving the auditory utterance within the window of silence to offset the synchrony of the auditory in relation to the visual clips, the combination of which formed the stimuli for the auditory-lead and auditory-lag stimuli.

Design and Procedure

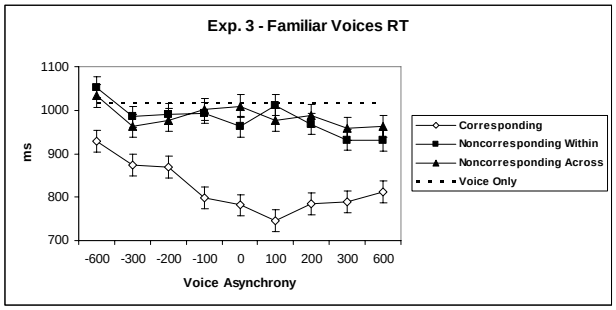
The experiment used the same instructions as the previous experiments, that participants should attentively view the visual stimuli, but should make familiar/unfamiliar responses exclusively based on the speaker's *voice*. Twenty-eight conditions of audiovisual stimulation were presented for both familiar and unfamiliar voices, comprising 12 trials each, resulting in 28 x 2 x 12 = 672 experimental trials, which were presented in randomized order in one randomised block of 672 trials. Breaks were allowed every 96 trials. In order to acquaint the participants to the task, the experimental trials were preceded by 30 practice trials during which each experimental condition was represented at least once.

The twenty-eight audiovisual conditions consisted of 9 asynchrony conditions in the *corresponding*, *noncorresponding-within* and *noncorresponding-across* conditions, plus voice only baseline. The asynchrony conditions consisted of 1) synchronous (0ms asynchrony), 2-5) 600, 300, 200 and 100 milliseconds auditory-lead (henceforth referred to as, -600ms, -300ms, -200ms, and -100ms) and 6-9) 600, 300, 200, and 100 milliseconds auditory-lag (henceforth referred to as, +600ms, +300ms, +200ms, and +100ms). Responses were measured relative to the onset of the auditory utterance.

Results

Reaction Time (RT) Data

a.



b.

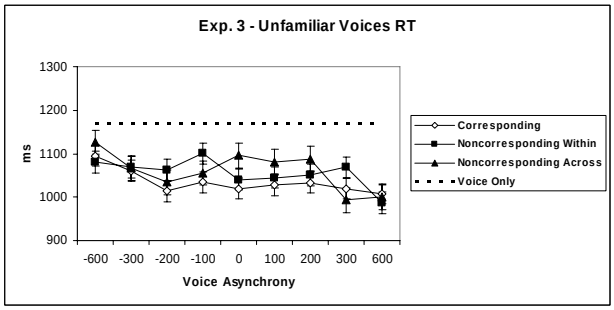


Figure 12a,b Mean reaction-times (RT) for familiar and unfamiliar voices.

Figure 12*a,b* display the mean RT data for Experiment 3. Mean RT data were submitted to an analysis of variance (ANOVA) with repeated measures for presentation condition (28 levels = (9 synchrony levels × 3 correspondence levels) + voice-only baseline) and familiarity.

For the RT data, a significant main effect was found for familiarity, $F(1, 23) = 30.59$, $p < 0.01$, which reflected that responses were generally faster for familiar voices compared to unfamiliar voices. There was also a significant main effect of presentation condition, $F(27, 621) = 12.48$, $p < 0.001$, and the effects of familiarity and presentation condition were moderated by a significant interaction, $F(27, 621) = 5.60$, $p < 0.001$. For familiar and unfamiliar voices only, there were main effects of presentation condition, $F(27, 621) = 16.59$, $p < 0.001$, and, $F(27, 621) = 3.09$, $p < 0.001$, respectively.

Once again, to more clearly ascertain the effects of the audiovisual conditions in comparison to the voice-only baseline condition, separate analyses were conducted comparing each condition to baseline performance.

Familiar Voices

Corresponding – The synchronised condition resulted in significant RT benefits in comparision to baseline performance, $F(1, 23) = 48.62$, $p < 0.001$. The -600ms condition displayed a nonsignificant trend for benefits, $F(1, 23) = 4.04$, $p = 0.056$, while -300ms, $F(1, 23) = 11.37$, $p < 0.01$, -200ms, $F(1, 23) = 14.41$, $p < 0.001$, and -100ms, $F(1, 23) = 55.08$, $p < 0.001$, demonstrated significant RT benefits in comparison to the voice-only baseline. For the auditory-lag stimuli, +600ms, $F(1, 23) = 26.55$, $p < 0.001$, +300ms, $F(1, 23) = 66.90$, $p < 0.001$, +200ms, $F(1, 23) = 36.31$, $p < 0.001$, and, +100ms $F(1, 23) = 55.80$, $p < 0.001$, also demonstrated significant RT benefits compared to voice-only baseline performance.

Noncorresponding-within – RTs in the synchronised condition were not significantly different from voice-only performance, $F(1, 23) = 2.18$, $p > 0.05$. None of the auditory-lead conditions differed significantly from baseline performance, while in the auditory-lag condition, the +600ms, $F(1, 23) = 7.58$, $p < 0.05$, and +300ms, $F(1, 23) = 5.00$, $p < 0.05$, conditions were the only ones to display RT benefits compared to baseline.

Noncorresponding-across – RTs in the synchronised condition were not significantly different from voice-only performance, $F(1, 23) = 0.02$, $p > 0.05$. None of the auditory-lead or auditory-lag conditions resulted in performance that was significantly different from baseline.

Unfamiliar Voices

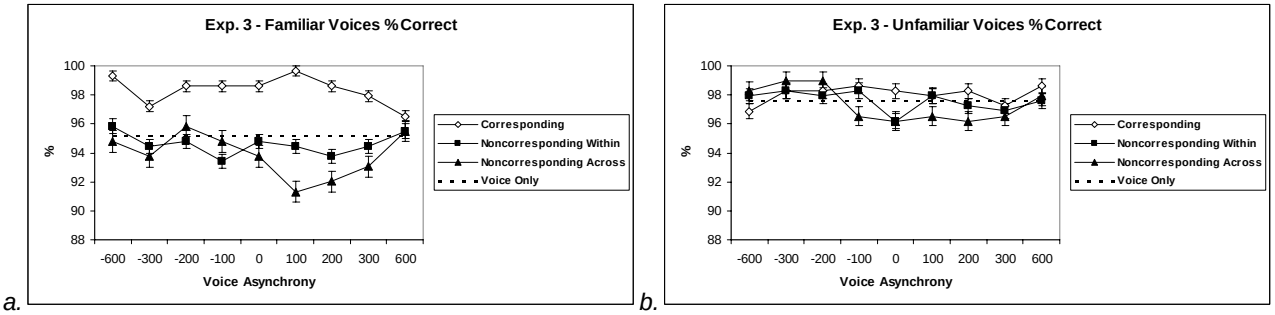
Corresponding – The synchronised condition resulted in significant RT benefits in comparison to baseline, $F(1, 23) = 20.95$, $p < 0.001$. The -600ms condition displayed a nonsignificant trend for benefits, $F(1, 23) = 4.19$, $p = 0.052$, while the -300ms, $F(1, 23) = 11.26$, $p < 0.01$, -200ms, $F(1, 23) = 16.19$, $p < 0.001$, and -100ms, $F(1, 23) = 8.74$, $p < 0.01$, conditions displayed significant RT benefits compared to the voice-only baseline. For the auditory-lag conditions, +600ms, $F(1, 23) = 18.29$, $p < 0.001$, +300ms, $F(1, 23) = 18.35$, $p < 0.001$, +200ms, $F(1, 23) = 12.19$, $p < 0.01$, and +100ms, $F(1, 23) = 13.21$, $p < 0.01$, conditions, all demonstrated significant RT benefits compared to the voice-only baseline.

Noncorresponding-within – The synchronised condition displayed significant RT benefits compared to baseline, $F(1, 23) = 9.11$, $p < 0.01$. The -600ms, $F(1, 23) = 6.13$, $p < 0.05$, -300ms, $F(1, 23) = 7.20$, $p < 0.05$, -200ms, $F(1, 23) = 6.97$, $p < 0.01$, conditions demonstrated significant RT benefits, while the -100ms, $F(1, 23) = 3.38$, $p = 0.079$, showed a nonsignificant trend for benefits compared to baseline performance. The +600ms, $F(1, 23) = 21.39$, $p < 0.001$, +300ms, $F(1, 23) = 6.67$, $p < 0.05$, +200ms, $F(1, 23) = 15.71$, $p < 0.001$, and +100ms, $F(1, 23) = 16.46$, $p < 0.001$, conditions demonstrated significant benefits compared to the voice-only baseline.

Noncorresponding-across – RTs in the synchronised condition did not differ significantly from baseline, $F(1, 23) = 2.73$, $p > 0.05$. The -600ms, $F(1, 23) = 2.25$, $p > 0.05$, was not significantly different, while the -300ms, $F(1, 23) = 10.78$, $p < 0.01$, -200ms, $F(1, 23) = 8.77$, $p < 0.01$ and -100ms, $F(1, 23) = 9.61$, $p < 0.01$, again demonstrated significant benefits compared to voice-only performance. The +600ms, $F(1, 23) = 37.36$, $p < 0.001$, and +300ms, $F(1, 23) = 24.73$, $p < 0.01$, conditions also showed significant RT benefits

compared to voice-only. The +200ms, $F(1, 23) = 4.24$, $p = 0.051$, and +100ms, $F(1, 23) = 3.56$, $p = 0.072$, displayed only nonsignificant trends for benefits in comparison to baseline.

Percentage correct data



a. *b.*

Figure 13a,b Mean percentage-correct data for familiar and unfamiliar voices.

Figures 13*a,b* display the mean percentage correct data for Experiment 3. An analogous ANOVA was performed on the percentage correct data. This revealed significant main effects of familiarity, $F(1, 23) = 5.98$, $p < 0.05$, and presentation condition, $F(27, 621) = 3.09$, $p < 0.001$, and that these effects were again moderated by a significant interaction, $F(27, 621) = 2.05$, $p < 0.01$. This indicated that presentation condition had a larger effect on familiar voices than unfamiliar voices. There was a significant main effects of presentation condition for familiar, $F(27, 621) = 3.55$, $p < 0.001$, but not for unfamiliar voices, $F(27, 621) = 0.95$, $p > 0.05$.

Familiar Voices

Corresponding – The synchronised condition resulted in significantly large benefits in comparison to baseline, $F(1, 23) = 8.10$, $p < 0.01$. The -600ms, $F(1, 23) = 13.77$, $p < 0.01$, condition also resulted in significant benefits to accuracy compared to baseline. The -300ms, $F(1, 23) = 1.67$, $p > 0.05$, condition was not significantly different from baseline, but the -200ms, $F(1, 23) = 6.00$, $p < 0.05$, and -100ms, $F(1, 23) = 6.93$, $p < 0.05$, demonstrated significant benefits to accuracy in comparison to baseline performance. The +600ms, $F(1, 23) = 1.00$, $p > 0.05$, condition was not significantly different from baseline, but the +300ms, $F(1, 23) = 3.99$, $p = 0.058$, condition demonstrated a nonsignificant trend for benefits compared to baseline performance. The +200ms, $F(1, 23) = 6.93$, $p < 0.05$, and +100ms, $F(1, 23) = 16.22$, $p < 0.001$, conditions demonstrated significant benefits compared to the voice-only baseline.

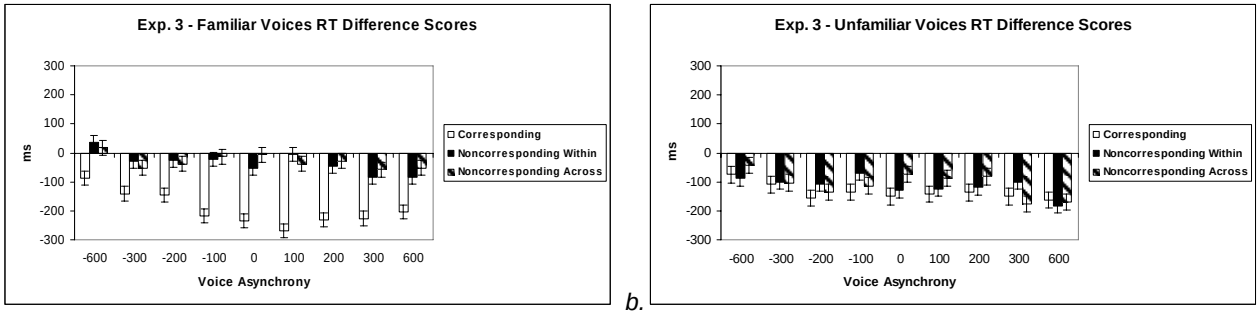
Noncorresponding-within – The synchronised condition did not differ significantly from baseline, $F(1, 23) = 0.05$, $p > 0.05$. Furthermore, none of the auditory-lead or auditory-lag conditions differed significantly from the baseline.

Noncorresponding-across – The synchronised condition also did not differ significantly from the voice-only baseline, $F(1, 23) = 0.66$, $p > 0.05$. None of the auditory-lead conditions differed significantly from baseline. For the auditory-lag conditions, +200ms, $F(1, 23) = 3.59$, $p = 0.071$, demonstrated a nonsignificant trend for costs, while the +100ms, $F(1, 23) = 4.48$, $p < 0.05$, condition was the only condition to display significant costs to accuracy in comparison to baseline performance.

Unfamiliar Voices

None of the percentage-correct data for the audiovisual conditions were significantly different from the voice-only baseline.

RT Difference Scores



a. *b.*

Figure 14a,b RT difference scores for familiar and unfamiliar voices.

As before, difference scores were calculated by subtracting the voice only baseline from all of the conditions, in order to more closely analyse the effects of the presentation conditions, specifically, evaluating the differences between the synchronous conditions and the asynchronous conditions at each level of correspondence – in relation to the voice only baseline. ANOVAs on these data involved repeated measures on voice asynchrony (9 levels, *synchronous, -600, -300, -200, -100, +600, +300, +200, and +100 millisecond voice asynchronies*), and face correspondence (*corresponding*, *noncorresponding-within*, and *noncorresponding-across*).

Familiar voices

Fig. 14*a* displays the RT Differences scores for familiar voice stimuli in Experiment 3. The ANOVA on these data revealed a significant main effect for correspondence, $F(2, 46) = 105.66$, $p < 0.001$, and presentation condition, $F(8, 184) = 7.07$, $p < 0.001$, and a significant interaction for correspondence by presentation condition, $F(16, 368) = 2.59$, $p < 0.01$. The *corresponding* condition displayed significant RT benefits compared to the *noncorresponding-within*, $F(1, 23) = 172.30$, $p < 0.001$, and *noncorresponding-across*, $F(1, 23) = 152.31$, $p < 0.001$, conditions. The comparison between the two *noncorresponding* conditions was not significant, $F(1, 23) = 0.14$, $p > 0.05$.

For *corresponding* stimuli, RT benefits for synchronous were significantly large compared to the -600ms, $F(1, 23) = 22.54$, $p < 0.001$, -300ms, $F(1, 23) = 9.82$, $p < 0.01$, and -200ms conditions, $F(1, 23) = 16.42$, $p < 0.001$, while the comparison of synchronous and -100ms was not significant, $F(1, 23) = 0.76$, $p > 0.05$. By contrast, none of the auditory-lag conditions were significantly different from the synchronous condition for *corresponding* stimuli, however, the +100ms condition showed numerically the largest RT benefits.

In the *noncorresponding-within* condition, there were RT costs for -600ms compared to the synchronous condition, $F(1, 23) = 10.18$, $p < 0.01$, but there were no other significant comparisons for this level of correspondence. The *noncorresponding-across* condition showed significant costs for the synchronous compared to the -300ms, $F(1, 23) = 4.34$, $p < 0.05$, and +600ms conditions, $F(1, 23) = 4.44$, $p < 0.05$.

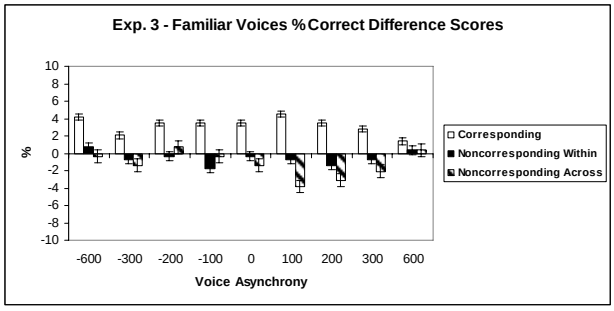
Unfamiliar Voices

Fig. 14*b* displays the RT Difference Scores for unfamiliar voices. The ANOVA for this data revealed a main effect of presentation condition, $F(8, 184) = 4.00$, $p < 0.001$, but not correspondence, $F(2, 46) = 2.21$, $p > 0.05$, with no significant interaction, $F(16, 368) = 1.32$, $p > 0.05$. The *corresponding* condition showed significant benefits to voice recognition performance compared to the *noncorresponding-across* condition, $F(1, 23) = 4.35$, $p < 0.05$, but there were no other significant comparisons between the correspondence conditions.

For *corresponding* stimuli, the synchronous condition showed RT benefits compared to the -600ms condition, $F(1, 23) = 4.69$, $p < 0.05$. None of the audio-lead and audio-lag conditions significantly differed from synchronous for the *corresponding* stimuli or *noncorresponding-within* stimuli. The noncorrespoding-across condition was similar, although the +600ms, $F(1, 23) = 7.58$, $p < 0.05$, and +300ms, $F(1, 23) = 14.06$, $p < 0.01$, condition showed significant RT benefits compared to the synchronous condition.

Percentage Correct Difference Scores

a.



b.

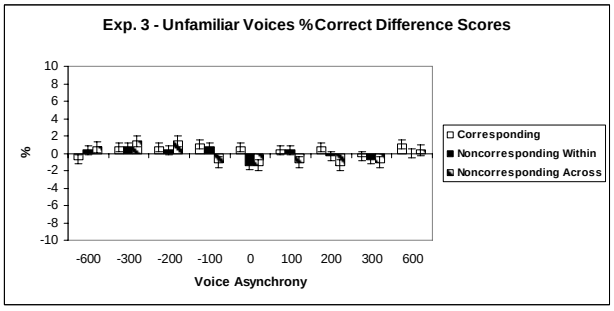


Figure 15a,b Percentage-correct difference scores for familiar and unfamiliar voices.

Familiar Voices

Figure 15*a* displays the percentage-correct difference scores for Experiment 3. The ANOVA on this data revealed main effects of correspondence, $F(2, 46) = 15.92$, $p < 0.001$, but not presentation condition, $F(8, 184) = 0.91$, $p > 0.05$, and no significant interaction, $F(16, 368) = 1.06$, $p > 0.05$. The *corresponding* condition showed accuracy benefits compared to the *noncorresponding-within*, $F(1, 23) = 25.47$, $p < 0.001$, and *noncorresponding-across*, $F(1, 23) = 22.47$, $p < 0.001$, conditions, but the *noncorresponding* conditions did not differ from each other, $F(1, 23) = 0.74$, $p < 0.001$. No significant differences existed between the synchronous and the asynchronous conditions for any of the three levels of stimulus correspondence, however, it seems that the benefits were numerically largest at +100ms.

Unfamiliar Voices

Figure 15*b* displays the percentage correct difference scores for unfamiliar voices in Experiment 3. The ANOVA on this data revealed no significant main effects and no interaction.

Discussion

Firstly, it is important to note that Experiment 3 displays particularly large facilitation effects for familiar *corresponding* stimuli compared to voice-only baseline performance. Performance was generally much faster and more accurate when *corresponding* stimuli were presented compared to the voice-only and *noncorresponding* conditions. However, it is also notable that, similar to Experiment 2, the effects for *noncorresponding* stimuli were not as strong as those seen in Experiment 1 and in Schweinberger et al. (2007). It seems again to be the case, that *noncorresponding* stimuli can be ignored when presented with a familiar voice. In this experiment however, this may be explained by the nature of the stimuli. Since all of the stimuli consist of static frames for the first 640ms, it may be the case that when the auditory stimulus is asynchronous to the visual stimulus – particularly for auditory-lead presentations – performance becomes similar to that observed previously for static stimuli. This possibility might be supported by the relatively poor performance accuracy for familiar *noncorresponding-across* stimuli between 0 and +300ms, while for auditory-lead presentations, performance is similar to baseline. The pattern for *noncorresponding* stimuli is generally unclear, and clarifying the results may require a larger sample of participants, since the number of conditions in the experiment limited the amount of repetitions that could be viably presented in each session.

Experiment 3 suggests that there is a degree of tolerance for asynchrony in the benefits and costs displayed by the previous experiments. Response accuracy was once again relatively high, overall higher than that seen in the previous experiments, and as mentioned previously, a ceiling effect may account for the relative lack of variance in the

accuracy data. The RT data for familiar voices showed again that the *corresponding* condition demonstrated significantly large RT benefits from baseline compared to the *noncorresponding* conditions, and this is clear from the data displayed in Figure 14*a*. For familiar *corresponding* audiovisual stimuli, there is the clearest indication of a temporal window for integration. There is a gradual trend of RT benefits as the asynchrony of the stimuli moves from -600 milliseconds to synchronous. The -600ms, -300ms, -200ms stimuli suffer significant RT costs in comparison to the synchronous condition, suggesting a gradual degradation of performance when the auditory speech precedes the visual speech by more than 100 milliseconds. There were no auditory-lag conditions that were significantly different from synchronous presentations, however it should be noted that +100ms showed numerically the largest benefits to RT performance.

The pattern for the percentage-correct data in the familiar *corresponding* condition is a little less clear, which may suggest that stimuli in this condition were too easily classified. Surprisingly, the -600ms condition showed high accuracy, when it may have been expected that it should be closer baseline performance. The -300ms condition displays the sort of accuracy that was expected, although it wasn't significantly different to the synchronous condition. The accuracy for the -200ms and -100 millisecond auditory-lead conditions were almost identical to the synchronous condition. This observation in some conditions may reflect a ceiling effect, that the task was simply too easy for asynchronies near to the synchronous condition.

With respect to the auditory-lag conditions, they demonstrated similar accuracies as seen for the synchronous condition, with a greater, but nonsignificant drop-off in performance in

the +600ms condition. These results may tentatively suggest that the temporal window for integration in the case of the familiar *corresponding* condition lies between -100 milliseconds and +300 milliseconds. This is larger, but qualitatively similar to the temporal window of integration suggested for the McGurk effect (van Wassenhove et al., 2007), although the precision of the estimation was much higher in that study than in the present experiment, as there was a considerably larger variety of asynchrony levels used.

The RT data for the familiar *noncorresponding* conditions generally displayed no discernable pattern, other than most of the asynchrony conditions were quite similar to the voice only baseline. Perhaps unexpectedly, the -600ms condition for both *noncorresponding* conditions displayed significantly slower responses than for the synchronous condition, where it may have been expected that more extreme asynchronies would result in performance closer to baseline. The percentage-correct data indicated the expected pattern for *noncorresponding-across* stimuli, where performance was gradually degraded as the modalities moved towards synchrony, however none of the asynchronous conditions were significantly different from the synchronous conditions.

It is notable that the +100ms condition displays numerically the largest benefits in the *corresponding* condition and the largest costs in the *noncorresponding-across* condition compared to the other levels of synchrony. These findings show striking similarities to the studies by (Munhall et al., 1996) and (van Wassenhove et al., 2007), where the McGurk illusions were slightly more likely to be observed for the +100ms condition compared to the synchronous condition. This may seem surprising, but may be explained by the nature of speech, where a small amount of auditory lag naturally occurs as sound waves travel

more slowly than visual signals (Massaro & Cohen, 1995). Furthermore, there is evidence from a study on “multimodal neurons” in the superior colliculus that response enhancements were optimal when multimodal stimuli are presented at 100-200 ms asynchronies. Such results suggest that multimodal mechanisms in the brain are tuned to the asynchronies most likely to occur.

The RT data for unfamiliar voices once again displayed no obvious pattern of significant effects. It may be worth noting however, for the *corresponding* and *noncorresponding-across* conditions, that between the suggested temporal window of -100 milliseconds to +300 milliseconds the RT data flattens somewhat, while showing costs and benefits for larger asynchronies. The accuracy data for the unfamiliar voices was particularly high with the -600ms stimuli in the *corresponding* condition showing the most costs to performance. While benefits and costs across asynchronies were quite erratic for the *noncorresponding-within* condition, the *noncorresponding* across condition showed a more obvious pattern, where costs increased from -300ms to synchronous, and the level of accuracy remained relatively similar to synchronous for the auditory lag conditions. This pattern may be more evident due to the presence of familiar faces. It could be argued that this condition is the most difficult of the experiment, as participants are presented with familiar faces with unfamiliar voices, so there is a response conflict caused by the differing familiarities of the stimuli, before the decision is made on the familiarity of the voice. Thus, the pattern may be more evident as it is not as severely affected by the ceiling effect seen for much of the previous accuracy data.

It seems possible then that the AVI shown previously can be achieved within the restrictions of a temporal window of integration similar to that suggested for the McGurk effect (van Wassenhove et al., 2007). The patterns shown particularly by the familiar *corresponding* and unfamiliar *noncorresponding-across* conditions suggest that there is something special about how familiar faces are integrated with voices. For these stimuli, the pattern of trends suggest a relatively large but qualitatively similar temporal window for person recognition as that demonstrated for speech recognition (van Wassenhove et al., 2007). This could be viewed as strong support for an AVI basis for the previously shown effects. However, despite the clear patterns of some of the data, the synchronous condition does not significantly differ from asynchronies outside of the suggested window enough, so any conclusions made on these data must be tentative at best. A much greater number of audiovisual conditions and longer stimuli presentations meant that the number of trials per condition had to be limited to prevent each experimental session from taking too long. Further research must use a larger sample to ascertain if the various nonsignificant trends could in fact be significant.

CHAPTER 4:

AUDIOVISUAL INTEGRATION DURING FACE RECOGNITION

EXPERIMENT 4

Introduction

Faces are more quickly and efficiently recognised than voices (Ellis, Jones, & Mosdell, 1997; Schweinberger et al., 1997), so it may stand to reason that visual stimuli should have strong effects on voice recognition performance. We have already established that AVI is presumably causing the effects seen in our previous experiments. The question remains however, whether it is possible that the auditory stimulus could exert some influence on face recognition. Although many audiovisual studies have shown effects where the visual stimulus alters the perception of the auditory stimulus, most famously in the ventriloquist and McGurk effects, evidence does exist for auditory stimuli altering visual perception (Shams et al., 2002; Shams, Kamitani, Thompson, & Shimojo, 2001). Although priming by voices can improve face recognition (Ellis, Jones, & Mosdell, 1997) there are few indications that voices have effects on face recognition during audiovisual stimulation. Audiovisual integration is usually dominated by the modality with the highest spatial resolution (Calvert et al., 1998). That is, when both modalities are clearly perceived, the modality with highest resolution – conveying the most detailed and reliable information - is vision, and this usually dominates auditory perception, as seen in the McGurk effect. It seems unlikely then, that voices should have a significant effect on face recognition. If there are indeed significant effects of voice identity on face recognition performance, it might suggest that multimodal representations of people are not completely visually-biased.

Method

Participants

Twenty two participants (20 females, mean age 20.9), all in regular contact with the lecturers used as familiar speakers completed the experiment. Participants were all undergraduates of the Friedrich-Schiller University Institute of Psychology, were offered a choice of money (5€ per hour) or course credit for their participation and filled out a questionnaire indicating their level of familiarity with the familiar and unfamiliar speakers.

Stimuli and Apparatus

Experiment 4 used the same apparatus as the first two experiments. Since the new task required subjects to identify whether the faces presented were familiar or unfamiliar, simple presentation of the clips would likely have led to ceiling performance, since faces are much more quickly and easily recognised than voices (Schweinberger, Herholz, & Sommer, 1997). Therefore, new visual stimuli were created using the forwards dynamic videos from the previous experiments as the source clips. The clips were altered to grey-scale and there was a linear blur-in during the first 1000 ms of each video. That is, the first frame of each video was blurred so as to be completely unrecognizable and linearly clarified until it was completely clear at the 1000 ms time-point. These alterations were used to decrease the rate at which perceptual information for face identification would become available and in an attempt to make the points in time at which a person could be recognised from the face more similar to the respective point in time for voices.

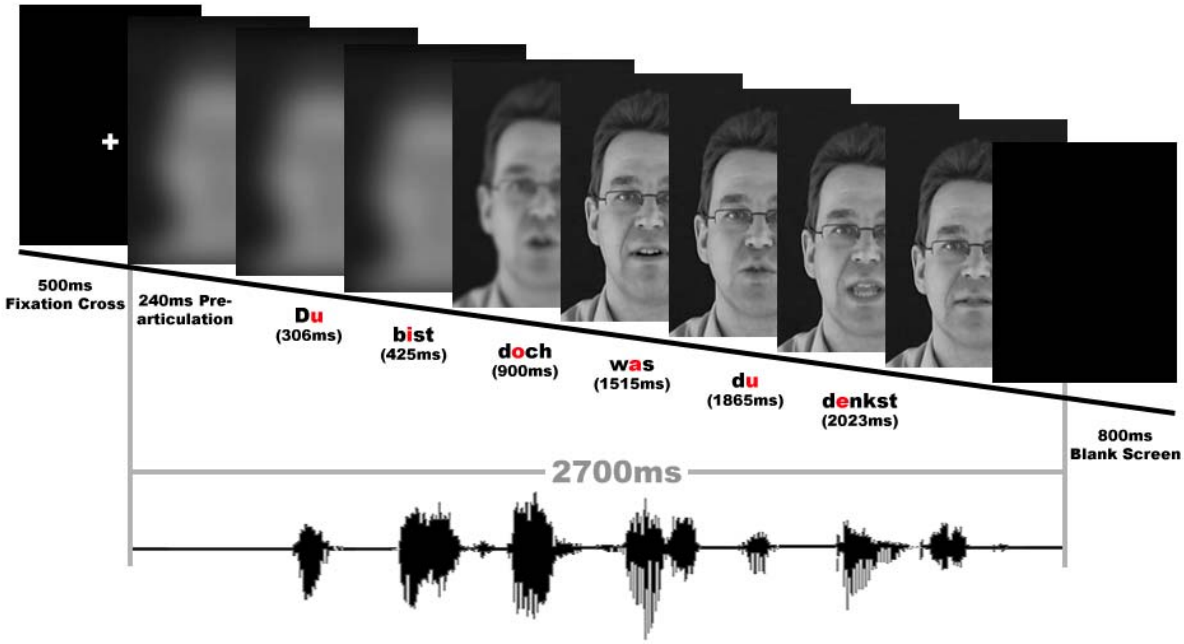


Figure 16 An example of a typical audiovisual trial in Experiment 4.

Stimuli were grey-scaled to rule out the influence of colour cues which may aid early face recognition. It has been shown that presentation of faces in grey-scale has no effect on audiovisual integration effects (Jordan, McCotter, & Thomas, 2000). The auditory clips remained the same as those used in the first two experiments.

Design and Procedure

Differing from our past two experiments, the instructions emphasized that participants should attend to the auditory stimuli, but should make familiar/unfamiliar responses exclusively based on the speaker's *face*. Once again, before each participant began the

experiment they were asked to complete a questionnaire in which they verified that they were highly familiar with the familiar speakers and were completely unfamiliar with the unfamiliar speakers. Four conditions of audiovisual stimulation were presented for both familiar and unfamiliar voices, comprising 36 trials each, resulting in 4 x 2 x 36 = 288 experimental trials, which were presented in randomized order in three consecutive blocks of 96 trials each. Breaks were allowed every 48 trials. In order to acquaint the participants to the task, the experimental trials were preceded by 16 practice trials during which each experimental condition was represented at least once.

The four audiovisual conditions were either 1) *face only* (no auditory stimulus), 2) *corresponding* (voice matched facial identity), 3) *noncorresponding-within* familiarity (for a familiar face, a familiar voice of differing identity was presented), 4) *noncorresponding-across* familiarity (for a familiar face, an unfamiliar voice was presented). The temporal attributes of the trials remained the same as the first two experiments and responses were measured in the same way.

Results

Reaction Time (RT) data

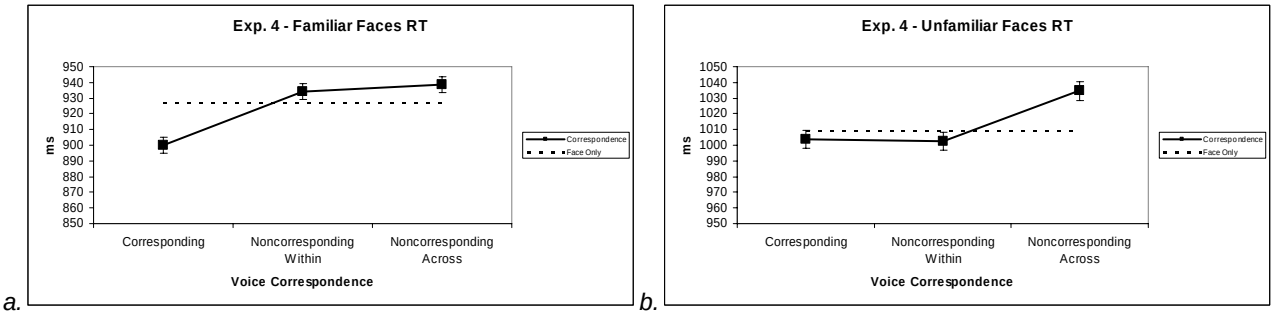


Figure 17a,b Mean reaction-time (RT) data for familiar and unfamiliar faces.

Figure 17*a,b* display the mean correct RT data for Experiment 4. As for the previous three experiments, the data were submitted to an analysis of variance (ANOVA), this time with repeated measures for presentation condition (4 levels) and face familiarity. Where appropriate, epsilon corrections for heterogeneity of covariances (Huynh et al., 1976) were performed throughout.

In the RT data, there were significant main effects of familiarity, $F(1, 19) = 66.62$, $p < 0.001$, reflecting faster responses to familiar than unfamiliar *faces*, and of presentation condition, $F(3, 57) = 10.70$, $p < 0.001$. These effects were moderated by an interaction of familiarity by presentation condition, $F(3, 57) = 5.19$, $p < 0.01$. This interaction reflected the fact that audiovisual presentation condition resulted in significantly faster RTs for familiar faces than for unfamiliar faces, although presentation condition was significant for both familiar *faces*, $F(3, 57) = 11.24$, $p < 0.001$, and unfamiliar *faces*, $F(3, 57) = 6.56$, $p < 0.001$.

Again, to more clearly evaluate the effects of the audiovisual conditions in comparison to the unimodal – in this case face-only – baseline condition, separate analyses were conducted comparing each condition to baseline performance.

Familiar faces

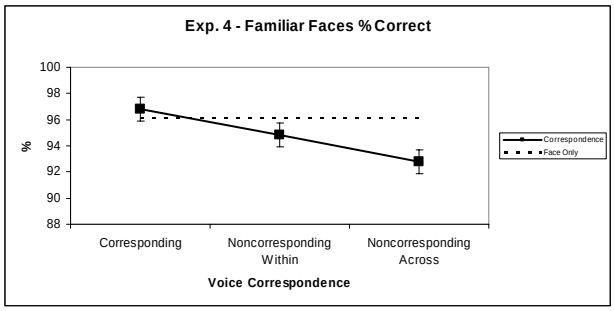
The *corresponding* condition displayed significantly large RT benefits in comparison to the face-only baseline, $F(1, 19) = 8.80$, $p < 0.01$. The *noncorresponding-within* condition was not significantly different from baseline, $F(1, 19) = 2.03$, $p > 0.05$, and the *noncorresponding-across* condition demonstrated a nonsignificant trend for RT costs compared to the face-only baseline, $F(1, 19) = 3.79$, $p = 0.067$.

Unfamiliar faces

The *corresponding* and *noncorresponding-within* conditions did not differ significantly from baseline, $Fs(1, 19) < 1$, $ps > 0.05$. However, the *noncorresponding-across* condition demonstrated significant RT costs compared to the face-only baseline, $F(1, 19) = 8.10$, $p < 0.05$.

Percentage-correct data

a.



b.

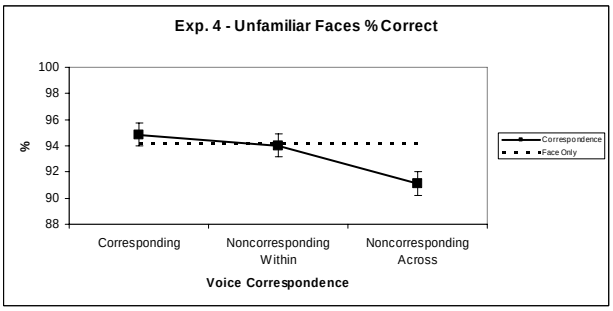


Figure 18a,b Mean percentage-correct data for familiar and unfamiliar faces.

Figure 18*a,b* displays the percentage-correct data for Experiment 4. The ANOVA revealed main effects of familiarity, $F(1, 19) = 6.01$, $p < 0.05$, and presentation condition, $F(3, 57) = 8.81$, $p < 0.001$. There was no significant interaction of familiarity by presentation condition, $F(3, 57) = 0.16$, $p > 0.05$. The effect of presentation condition on response accuracy was significant for both familiar faces, $F(3, 57) = 4.21$, $p < 0.01$, and for unfamiliar faces, $F(3, 57) = 3.56$, $p < 0.05$.

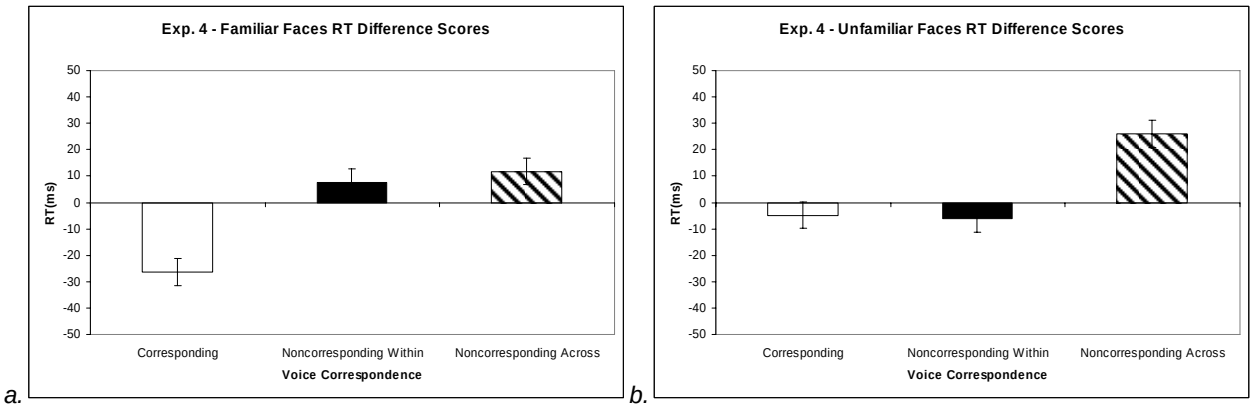
Familiar faces

The *corresponding* and *noncorresponding-within* conditions were not significantly different from baseline performance, Fs$(1, 19) < 1.6$, ps > 0.05. The *noncorresponding-across* condition showed significant costs to response accuracy compared to the face-only baseline, $F(1, 19) = 4.59$, $p < 0.05$.

Unfamiliar faces

Again, the *corresponding* and *noncorresponding-within* conditions did not significantly differ from face-only performance, *F*s(1, 19) < 1, *p*s > 0.05. Yet again, the *noncorresponding-across* condition demonstrated significant costs to response accuracy compared to the face-only baseline, $F(1, 19) = 5.62$, $p < 0.05$.

RT difference scores



a. *b.*

Figure 19a,b RT difference scores for familiar and unfamiliar faces.

Familiar faces

Figure 19*a* displays the RT difference scores for familiar faces in Experiment 4. The ANOVA on this data revealed a significant main effect of correspondence, $F(2, 38) = 15.46$, $p < 0.001$. The *corresponding* condition displayed significantly large benefits to face recognition compared to the small costs for the *noncorresponding-within*, $F(1, 19) = 18.81$, $p < 0.001$, and *noncorresponding-across* conditions,

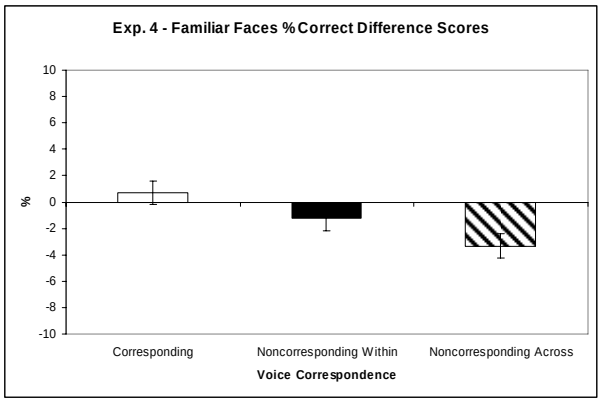
$F(1, 19) = 25.74$, $p < 0.001$. The two *noncorresponding* conditions did not differ significantly from each other, $F(1, 19) = 0.36$, $p > 0.05$.

Unfamiliar Faces

Figure 19*b* displays the RT difference scores for unfamiliar faces in Experiment 4. The ANOVA on this data showed a main effect of correspondence, $F(2, 38) = 11.54$, $p < 0.001$. The *corresponding* condition was not significantly different from the *noncorresponding-within* condition, $F(1, 19) = 0.03$, $p > 0.05$. The *noncorresponding-across* displayed costs to RTs compared to the *corresponding* condition, $F(1, 19) = 14.58$, $p < 0.01$. The *noncorresponding-across* condition demonstrated significantly large RT costs compared to the *noncorresponding-within* condition, $F(1, 19) = 21.22$, $p < 0.001$.

Percentage-correct difference scores

a.



b.

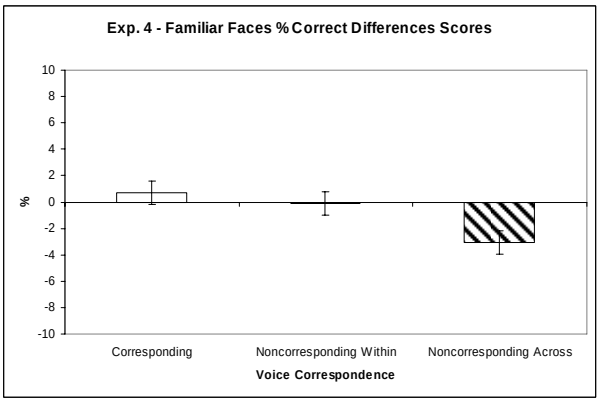


Figure 20a,b Percentage-correct difference scores for familiar and unfamiliar faces.

Familiar Faces

Figure 20a displays the percentage-correct difference scores for familiar faces in Experiment 4. The ANOVA for these data revealed a main effect of correspondence, $F(2, 38) = 5.79$, $p < 0.01$. The *corresponding* condition displayed a nonsignificant trend for performance benefits compared to the *noncorresponding-within* condition, $F(1, 19) = 3.71$, $p = 0.069$. Importantly, the *noncorresponding-across* condition, $F(1, 19) = 11.93$, $p < 0.01$, displayed significantly large costs to performance compared to the *corresponding* condition. The difference between the *noncorresponding-within* and *noncorresponding-across* conditions was not significant, $F(1, 19) = 2.38$, $p > 0.05$.

Unfamiliar faces

Figure 20*b* displays the percentage-correct data for unfamiliar faces in Experiment 4. There was a significant main effect of correspondence, $F(2, 38) = 4.63$, $p < 0.05$. The difference between *corresponding* and *noncorresponding-within* was not significant, $F(1, 19) = 0.38$, $p > 0.05$. The *noncorresponding-across* condition displayed significantly large costs to face recognition performance compared to the *corresponding* condition, $F(1, 19) = 9.55$, $p < 0.01$. The *noncorresponding-across* condition also suffered significantly larger costs in accuracy compared to the *noncorresponding-within* condition, $F(1, 19) = 4.93$, $p < 0.05$.

Discussion

Experiment 4, perhaps surprisingly, shows significant effects of voice stimuli on face recognition performance. It seems that, as was apparent for voices, when a familiar *corresponding* audiovisual pairing is presented, RTs were significantly faster than the other conditions. The *noncorresponding* conditions showed slight costs to face recognition although these were not significant. There were also slight accuracy benefits for familiar *corresponding* faces, while the *noncorresponding* conditions represented costs, where the *noncorresponding-across* condition displayed largest costs to face recognition accuracy. The RT data for unfamiliar face recognition was similar to baseline for the *corresponding* and *noncorresponding-within* condition. The *noncorresponding-across* condition however, showed significant costs to response time compared to the other conditions. Similar was true for the accuracy data, where apparently *noncorresponding-across* stimuli resulted in relatively large costs to face recognition accuracy in comparison to the *corresponding* and *noncorresponding-within* conditions.

These results can perhaps be seen as a departure from previous research on auditory effects during face recognition (Joassin, Maurage, Bruyer, Crommelinck, & Campanella, 2004). Joassin et al. (2004) found that RTs were slower for audiovisual stimulus presentations compared to face-only presentations. The results of the current experiment, however, most notably demonstrate slight but significant benefits to face recognition when a voice of *corresponding* identity is presented. Although the degradation of the visual

stimulus did as was aimed, by delaying the recognition of the visual stimulus until the voice stimulus could be recognised, it is unlikely that the degradation confounds the present effects. The findings of significant benefits for *corresponding* stimuli and significant costs for *noncorresponding-within* stimuli, suggests that voice influences face recognition in this case. The speed of reaction times and the high accuracy of performance suggests that the voices interact with face recognition before the face is entirely clear. It may be the case that, in the absence of colour cues and fine-detail, participants were able to perceive some of the idiosyncratic facial movements in making their decision. Since the identifying properties of the voice shares some attributes with the coarse movements, early integration may occur and facilitate face recognition performance. Where the voice is of a different identity, similar integrative processes may begin upon the gradual clarity of the the facial movements. Since the voice is synchronised with the facial movements, but is not representative of the idiosyncratic movements seen, it may cause a conflict in the subject, slowing reaction times and causing errors. Any conclusions based on this data however, must be tentative in comparison with the first three experiments, since the actual effects observed are of a much smaller magnitude than those seen for voice recognition performance.

CHAPTER 5:

GENERAL DISCUSSION AND OUTLOOK

General Discussion and Outlook

These four experiments sought to investigate the effects of AVI during identity perception and provide indications that multimodal representations of familiar people may be encoded in long-term memory. In general, the results of the previous experiments strongly suggest that AVI has a role to play in person perception. Much of the research into AVI has concentrated on speech perception and it is well established that the integration of the two modalities contributes to the perception of speech as it is seen and heard. As mentioned in Chapter 1, while the McGurk illusion is accepted as evidence of AVI in speech perception (see (Colin & Radeau, 2003; McGurk et al., 1976), for a review) and is often used as a tool for investigating the properties of AVI (Sekiyama, 1997; Brancazio & Miller, 2005; Pare, Richler, & Ten Hove, 2003), evidence of integrative effects during identity perception has been relatively sparse and preliminary until now (Campanella et al., 2007). Person perception research has often concentrated on a single modality or used static face presentations in showing crossmodal identification effects. In the cases where studies have used dynamic stimuli to demonstrate audiovisual effects for identity recognition, they are often sequential matching tasks (Kamachi et al., 2003; Lachs et al., 2004a). At the time of writing, few experiments have been published which demonstrate benefits and costs to person perception during audiovisual stimulation. One study that employed audiovisual presentations demonstrated the effects of audiovisual presentation on the recognition of previously learned face-voice pairs (Joassin et al., 2004), and found intermediate performance for audiovisual stimuli compared to face (fastest) and voice stimuli (slowest). The first direct evidence of the effects of various

conditions of audiovisual presentation was provided by our group (Schweinberger et al., 2007) and the experiments in Chapter 2 helped to strengthen the findings from that study.

Experiment 1 demonstrated that patterns of benefits and costs to familiar voice recognition, regardless of the identity of the face presented, were larger for the synchronised dynamic presentations compared to the static presentations. Voice recognition was fastest and most accurate when presented with a dynamic presentation of a face of *corresponding* identity. Performance suffered most when a familiar voice was presented with an unfamiliar moving face, and was also negatively affected by a *noncorresponding* familiar face, while static pictures had little effect on voice recognition performance. These data are largely in line with previous findings using similar methods (Schweinberger et al., 2007), with perhaps the notable exception that significant costs were incurred when a familiar voice was presented with a familiar face of a different identity. The finding that performance improved for unfamiliar voices in Experiment 1 for any visual stimulus, are more difficult to interpret, but may signify the familiarisation of unfamiliar voices over the course of the experiment. If this is indeed the reason why nearly all of the conditions in which a visual stimulus was presented showed benefits in recognition of a voice as unfamiliar, it makes the small costs caused by the presentation of a familiar face with an unfamiliar voice all the more striking. It seems more likely however, that in situations of perceptual uncertainty, the unfamiliar faces may help as a cue for correctly responding “unfamiliar” to an unfamiliar voice.

Static faces did not greatly affect performance when they were known to be of a different identity to the voice. This might suggest that in this context, static faces can be ignored, while dynamic faces, in most cases apparently cannot be ignored, and cause slower RTs and poorer response accuracy. These findings strongly suggests that AVI is a relevant factor to identity recognition. It is likely that the relatively small facilitation effects shown for static faces of *corresponding* identity to the voice, improve performance by acting as a cue to voice identity, similar to the effects of short-term crossmodal priming (Ellis et al., 1997). The effects of dynamic presentations are considered to demonstrate AVI, although the possibility of dynamic stimuli simply containing more identity information than static faces (Lander et al., 2005), could not be completely ruled out due to the design of Experiment 1. Experiment 2, therefore, sought to investigate if facial movement alone was the basis for the apparent AVI effects.

Previous research by Lander and colleagues has indicated that dynamic faces are easier to recognise (Lander et al., 2005; Lander et al., 2000; Lander, Christie, & Bruce, 1999) while other research shows that face matching is improved when dynamic stimuli are used (Thornton & Kourtzi, 2002). As mentioned in Chapter 2, it is conceivable that the identity information in moving faces, absent from static pictures, may cause the effects found previously. Experiment 2 tested this hypothesis by replacing the static condition with a backwards-dynamic video condition. Although the pattern was less clear, the results indicate that in some cases forwards and backwards videos are different in their effects, particularly for familiar voices paired with a face of *corresponding* identity. The results from Experiment 2 can be seen as suggesting that the extra information-content in

dynamic compared to static presentations, cannot completely account for the dynamic-static differences found previously, but they cannot rule out the possibility that the extra information-content contributes to the effects. However, I believe that there is a lack of significant comparisons between the forwards and backwards data because the conditions are too similar: The backwards stimuli were created to manufacture asynchrony between the voice and face stimuli while still retaining the dynamic qualities of the speaker's face. It is possible that, even when reversed, the backwards facial movements are in approximate synchrony with the voice and are unable to be completely ignored in the same way as static faces. This may account for the similarities in performance between forwards and backwards stimuli. All in all, Chapter 2 gives some strong indications of AVI in identity recognition, but cannot disprove the suggestion that part of the difference between the data for static and dynamic stimuli is attributable to the respective amounts of identity information available in the two types of stimuli.

Studies on the McGurk illusion have suggested that substantial asynchronies between the modalities can still result in the fusion effects (Munhall et al., 1996; van Wassenhove et al., 2007). Munhall et al. (1996) demonstrated that McGurk responses were still prevalent within a temporal window of 60ms auditory-lead to 240ms auditory-lag, while van Wassenhove et al. (2007) more recently demonstrated a temporal window of integration between 30 milliseconds auditory-lead to 170 milliseconds auditory-lag. This temporal window for integration may help to explain why many of the forwards versus backwards comparisions in Experiment 2 were not significant. Due to the short durations between the consonant/vowel onsets presented in the video stimuli, reversing the videos would not

necessarily have manufactured systematic asynchrony which was significantly different to the forwards video stimuli. Experiment 3 was conducted on this basis, in an attempt to perform a more controlled manipulation of audiovisual asynchrony. The RT data for familiar voices, when the face matched the voice, was particularly striking as it seemed to follow a similar pattern as the data found for the McGurk illusion in van Wassenhove et al. (2007). Taking the nonsignificant trends into account for synchronous versus asynchronous comparisons, the data leads to the suggestion that there exists a temporal window of integration for familiar identity perception between 100 milliseconds auditory-lead and 300 milliseconds auditory-lag.

A general degradation of response accuracy was observed for familiar voices presented with unfamiliar faces as the asynchronies moved from auditory-lead to synchronous. Costs were greatest between 0 milliseconds asynchrony (synchronous) and 300 milliseconds auditory-lag, though once again, any conclusions made on this data can only be tentative as only the 300 millisecond auditory-lead condition differed significantly from the synchronous condition. The tentative suggestion in this case, is that when the stimuli are of different identities, they are less likely to have an effect on voice recognition as the asynchrony between the modalities increases. To provide a more exact measure of the temporal window of integration for the present effects, and to lend extra weight to the tentative suggestions made the nonsignificant trends observed, will require increasing the power of the design. Further studies may also wish to investigate asynchrony using better temporal resolution to give a more exact measure of the limits of this time window. It should also be noted that the response accuracy data indicated that accuracy was high

across all conditions and there were few asynchronies that significantly differed from synchronous. While several AVI experiments have used degraded stimuli to show benefits of integration (Lachs et al., 2004b; Rosenblum, Johnson, & Saldana, 1996), the experiments here used stimuli that were as clear as possible. This is a possible reason for the lack of obvious effects in much of the accuracy data. The clarity of the stimuli may have made the task too easy for the participants, resulting in a ceiling-effect.

Another possible reason that significant effects were not apparent, is that repeated asynchrony can lead to a shift in perceived synchrony (Navarra et al., 2005; Vatakis, Navarra, Soto-Faraco, & Spence, 2007). As the stimuli were randomised, it is possible that at times, preceding stimuli repetitions could have resulted in altered perception of some stimuli, causing unexpected benefits and costs to performance. Such benefits and costs may have combined under some asynchronies to cancel out any possible effects on accuracy. Due to a lack of significant comparisons between the synchronous condition and the asynchronous conditions in Experiment 3, the overriding implication is that the familiar *corresponding* stimuli are perceived or processed in a different way from the other conditions. Experiment 3 gives indications that AVI in identity recognition may be similarly flexible with regards to synchrony as has already been shown for the McGurk effect (van Wassenhove et al., 2007; Munhall et al., 1996). However, in order to be able to draw firmer conclusions on a possible temporal window of integration, a paradigm allowing for more repetitions of each stimulus needs to be tested. Chapters 2 and 3 certainly give indications that AVI is an important part of familiar person identification.

The benefits and costs of pairing voices with familiar faces has been a running theme throughout these studies. Voice recognition benefits occurred when familiar faces were presented in synchrony with familiar voices and costs were incurred when familiar faces were presented with unfamiliar voices. It is accepted that in audiovisual presentations, the visual modality dominates as it has the highest degree of spatial resolution (Calvert et al., 1998). The McGurk illusion is a prime example of the influence an incongruent visual stimulus has on the perception of an auditory stimulus. Since familiar faces are highly salient stimuli, it is considered unlikely that recognition of familiar faces should be affected by auditory stimuli. Priming studies (Buelthoff and Newell, 2004) provide evidence using previously learned face-voice pairs, that face recognition performance can be improved when preceded by the *corresponding* voice prime. Experiment 4 had the intention of testing whether face recognition could be influenced by different voice presentations. The findings were surprising in the light of data submitted by (Joassin et al., 2004), who suggested that while audiovisual presentations were faster and more accurate than voice-only presentations, face-only was fastest and most accurate. Experiment 4 however, found small but significant benefits for familiar face recognition when faces were presented with a voice of *corresponding* identity, and similarly small but significant costs to face recognition when familiar faces were paired with unfamiliar voices. Costs were also observed for unfamiliar faces presented with familiar voices. It may be the case that when colour cues and fine detail are absent, the face movements become influential at a point before the face is fully clear to the participants, so the integration of the voice with the characteristic movements could be the start of integrative processing, which leads to the respective benefits and costs.

It should be noted though, that in some cases, especially in Experiments 2 and 3, the effects of *noncorresponding* stimuli were not as consistent as those seen in the other experiments. Further research in this area perhaps requires a larger sample of stimuli in order to clearly show the effects of *noncorresponding* stimuli on voice recognition performance. The current stimuli were chosen because the four familiar speakers had regular contact with the majority of psychology students at Freidrich-Schiller-University, Jena. A possibility for widening the stimuli pool and the sample of participants, could be something in similar to Sheffert et al., (2004), where familiarity was manufactured through training sessions. It may be interesting to investigate whether the effects observed in the previous experiments can also be achieved with newly-learned stimuli, however it could be argued that such learned-familiarity stimuli are not necessarily representative of personally familiar people. It should also be a priority for future research on this theme to clarify the effects of *noncorresponding* stimuli further. It would appear that the general effects of *noncorresponding* stimuli in Experiments 1 and 4 conflict with the findings of Experiments 2 and 3 to a certain extent, so the nature of AVI in the case of *noncorresponding* identity is particularly worthy of further investigation.

The relatively persistent finding, however, that systematic benefits and costs to recognition performance occur with the presentation of familiar stimuli, suggests that the properties of AVI for familiar people are different from the processing of unfamiliar stimuli. This leads to the suggestion that multimodal representations of familiar people may be stored in long-term memory. Chapter 2 and 3 suggest that synchronised familiar faces of *corresponding* identity to the voice, result in strong recognition benefits in comparision to

the voice-only condition. Previous research into recognition memory for familiar voices has, most often, used unimodal stimuli (eg. (Schweinberger, Herholz, & Stief, 1997; Schacter & Church, 1992)). The vast majority of research into face recognition memory has involved the presentation of static face stimuli (Bruce & Valentine, 1985; Ellis, Shepherd, & Davies, 1979). As mentioned earlier, some research has been conducted on familiar faces using moving stimuli (Lander et al., 2005; Lander & Bruce, 2004; Lander et al., 2000), finding a benefit to recognition when dynamic faces are presented. Furthermore, it has been suggested that facial movement may become more important as experience with a face increases (O'Toole, Roark, & Abdi, 2002). These findings perhaps carry the implication that the typical facial movements associated with familiar people may be stored in long-term memory along with the facial identity. The possible existence of facial motion schemata in long-term memory for familiar people may contribute greatly to the AVI witnessed here. Since the current experiments show that the clearest effects for audiovisual pairings in which a moving familiar face is present, it is conceivable that the representation of a familiar person in memory is more easily accessed by dynamic stimulation, and it affects performance accordingly (Lander et al., 2005). Typical expressions associated with familiar faces have also been suggested as being encoded in face memory (Kaufmann & Schweinberger, 2004), suggesting that long-term memory preserves more than just a static facial representation of familiar people. The findings observed in Experiment 4 may take this idea further, it may be the case that there is a multimodal aspect to the encoding of familiar people in long-term memory. The results of that experiment suggest that face recognition is improved by *corresponding* audiovisual stimulation, leading to the possibility that representations of people in memory can be more quickly and easily accessed under such circumstances.

The implication that memory representations for familiar people may be more elaborate than is often tested may allow for the possibility that multimodal person representations exist in long-term memory. Experiment 4 demonstrated small but significant benefits and costs to familiar face recognition when familiar and unfamiliar voices were presented. Brain imaging studies, using unimodal stimuli, have demonstrated that areas of the brain normally associated with audition are activated during face perception (van Wassenhove et al., 2005), while other studies have demonstrated that areas most often associated with face recognition are activated when speech is heard (von Kriegstein, Kleinschmidt, Sterzer, & Giraud, 2005). Von Kriegstein et al., (2005) found that in a task emphasizing speaker recognition, the fusiform gyrus was activated when participants heard a voice. They go on to suggest that AVI is unlikely to occur in “supramodal” areas (such as the SC as posited by (Stein et al., 1993), but that the auditory and visual areas may share information about person identity. These findings may form the basis for audiovisual encoding of people in memory.

In the context of the present effects however, such early perceptual activation of crossmodal areas may imply that person representations are indeed encoded in a multimodal fashion. The suggestion from these brain imaging studies is that audiovisual stimuli do not necessarily depend on a cortical integration centre of convergence, but that the auditory and visual areas are able to directly interact with each other in perceiving audiovisual stimuli. Evidence that audiovisual input can improve voice learning (Sheffert &

Olson, 2004) and that familiar face-voice priming can result in recognition benefits (Schweinberger et al., 1997), might lend further weight to the suggestion that multimodal and dynamic person representations may exist in long-term memory.

Overall, the experiments give indications that audiovisual integration is an important aspect of person identification. They also suggest that audiovisual integration is particularly important for the recognition of familiar people, which suggests that multimodal representation of familiar people may be held in long-term memory. Future studies should attempt to extend these findings with brain imaging techniques. Functional resonance imaging studies have recently been used to provide data on how temporal asynchrony affects audiovisual processing (Noesselt et al., 2007). To extend the current data, it is important to find a way of clarifying the results of Experiment 3, perhaps by using a similar fine temporal resolution for stimulus asynchrony as that used by (van Wassenhove et al., 2007).

Uncovering the EEG and fMRI correlates of AVI in person perception would also help to give indications of the underlying processes. It might be interesting to investigate how audiovisual stimuli, in the case of a speaker identification task, differ from unimodal stimuli. The processing of famous faces has been shown to result in more widespread activation than for learned and unfamiliar faces (Leveroni et al., 2000). Therefore, it may be of future relevance to further investigate the effects seen in Experiment 1 using fMRI to investigate the differences in activation between dynamic and static audiovisual

presentations of familiar and unfamiliar people. Displaying such anatomical differences between these stimulus presentations may help provide further support for the suggestion that multimodal representations of familiar people exist in long-term memory.

Furthermore, it might be expected that the synchrony of the current stimuli, regardless of correspondence, would demonstrate early ERPs that would be similar across the correspondence conditions, due to the possibility that synchronous stimuli are automatically integrated (van Wassenhove et al., 2005). Therefore, investigating relatively late differences in the ERP signals for each condition of correspondence, might be more relevant in helping to understand the temporal aspects of facilitation and inhibition of performance during *corresponding* and *noncorresponding* audiovisual presentations. On the other hand, recent research has investigated very early effects during audiovisual stimulation, by analyzing gamma-band oscillations (Doesburg et al., 2008; Senkowski, Talsma, Grigutsch, Herrmann, & Woldorff, 2007). There appears to be a great amount of potential for the technique in further investigating the effect of audiovisual asynchrony, and it may be interesting to investigate early interactions in the context of audiovisual person perception. The indications from the four experiments in this thesis are that audiovisual integration plays a significant role in the recognition of familiar people and that long-term memory may be able to store dynamic multimodal representations of know people. There is still much to be investigated in order to gain a more detailed understanding of AVI in person recognition, but the current findings, and recent developments in the field certainly indicate that it is an important aspect of person perception which requires more in-depth analysis.

Reference List

Bernstein, L. E., Auer, E. T., Wagner, M., & Ponton, C. W. (2008). Spatiotemporal dynamics of audiovisual speech processing. *NeuroImage, 39,* 423-435.

Brancazio, L. & Miller, J. L. (2005). Use of visual information in speech perception: Evidence for a visual rate effect both with and without a McGurk effect. *Perception & Psychophysics, 67,* 759-769.

Bruce, V. (1990). Face recognition. In M.W.Eysenck (Ed.), *Cognitive Psychology. An international review* (pp. 221-263). Chichester, New York: Wiley.

Bruce, V. & Valentine, T. (1985). Identity priming in the recognition of familiar faces. *British Journal of Psychology, 76,* 373-383.

Bruce, V. & Valentine, T. (1988). When a nod's as good as a wink. The role of dynamic information in facial recognition. In M.M.Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues. Vol. 1: Memory in everyday life* (pp. 169-174). Chichester, New York: Wiley.

Calvert, G. A., Brammer, M. J., & Iversen, S. D. (1998). Crossmodal identification. *Trends in Cognitive Sciences, 2,* 247-253.

Campanella, S. & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences, 11,* 535-543.

Colin, C. & Radeau, M. (2003). The McGurk illusions in speech: 25 years of research. *Annee Psychologique, 103,* 497-542.

Doesburg, S. M., Emberson, L. L., Rahi, A., Cameron, D., & Ward, L. M. (2008). Asynchrony from synchrony: long-range gamma-band neural synchrony accompanies perception of audiovisual speech asynchrony. *Experimental Brain Research, 185,* 11-20.

Ellis, H. D., Jones, D. M., & Mosdell, N. (1997). Intra- and inter-modal repetition priming of familiar faces and voices. *British Journal of Psychology, 88,* 143-156.

Ellis, H. D., Shepherd, J. W., & Davies, G. M. (1979). Identification of familiar and unfamiliar faces from internal and external features: Some implications for theories of face recognition. *Perception, 8,* 431-439.

Howard, I. P. & Templeton, W. B. (1966). *Human Spatial Orientation*. London: Wiley.

Huynh, H. & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics, 1,* 69-82.

Joassin, F., Maurage, P., Bruyer, R., Crommelinck, M., & Campanella, S. (2004). When audition alters vision: an event-related potential study of the cross-modal interactions between faces and voices. *Neuroscience Letters, 369,* 132-137.

Jones, J. A. & Jarick, M. (2006). Multisensory integration of speech signals: the relationship between space and time. *Experimental Brain Research, 174,* 588-594.

Jordan, T. R., McCotter, M. V., & Thomas, S. M. (2000). Visual and audiovisual speech perception with color and gray-scale facial images. *Perception & Psychophysics, 62,* 1394-1404.

Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). 'Putting the face to the voice': Matching identity across modality. *Current Biology, 13,* 1709-1714.

Kaufmann, J. M. & Schweinberger, S. R. (2004). Expression influences the recognition of familiar faces. *Perception, 33,* 399-408.

Keetels, M. & Vroomen, J. (2008). Temporal recalibration to tactile-visual asynchronous stimuli. *Neuroscience Letters, 430,* 130-134.

Koppen, C. & Spence, C. (2007). Audiovisual asynchrony modulates the Colavita visual dominance effect. *Brain Research, 1186,* 224-232.

Lachs, L. & Pisoni, D. B. (2004a). Crossmodal source identification in speech perception. *Ecological Psychology, 16,* 159-187.

Lachs, L. & Pisoni, D. B. (2004b). Specification of cross-modal source information in isolated kinematic displays of speech. *Journal of the Acoustical Society of America, 116,* 507-518.

Lander, K. & Bruce, V. (2000). Recognizing famous faces: Exploring the benefits of facial motion. *Ecological Psychology, 12,* 259-272.

Lander, K. & Bruce, V. (2004). Repetition priming from moving faces. *Memory & Cognition, 32,* 640-647.

Lander, K., Christie, F., & Bruce, V. (1999). The role of movement in the recognition of famous faces. *Memory & Cognition, 27,* 974-985.

Lander, K. & Chuang, L. (2005). Why are moving faces easier to recognize? *Visual Cognition, 12,* 429-442.

Leveroni, C. L., Seidenberg, M., Mayer, A. R., Mead, L. A., Binder, J. R., & Rao, S. M. (2000). Neural systems underlying the recognition of familiar and newly learned faces. *The Journal of Neuroscience, 20,* 878-886.

Lewald, J., Ehrenstein, W. H., & Guski, R. (2001). Spatio-temporal constraints for auditory-visual integration. *Behavioural Brain Research, 121,* 69-79.

Lewald, J. & Guski, R. (2003). Cross-modal perceptual integration of spatially and temporany disparate auditory and visual stimuli. *Cognitive Brain Research, 16,* 468-478.

Maravita, A., Bolognini, N., Bricolo, E., Marzi, C. A., & Savazzi, S. (2008). Is audiovisual integration subserved by the superior colliculus in humans? *NeuroReport, 19,* 271-275.

Massaro, D. W. & Cohen, M. M. (1995). Perceiving talking faces. *Current Directions in Psychological Science, 4,* 104-109.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746-748.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics, 58,* 351-362.

Munhall, K. G. & Vatikiotis-Bateson, E. (2004). Spatial and temporal constraints on audiovisual speech perception. In G.Calvert, C. Spence, & B. E. Stein (Eds.), *The Handbook of Multisensory Processes* (pp. 177-188). Cambridge, Mass.: MIT Press.

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., & Spence, C. (2005). Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Cognitive Brain Research, 25,* 499-507.

Noesselt, T., Rieger, J. W., Schoenfeld, M. A., Kanowski, M., Hinrichs, H., Heinze, H. J. et al. (2007). Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *Journal of Neuroscience, 27,* 11431-11441.

O'Toole, A. J., Roark, D. A., & Abdi, H. (2002). Recognizing moving faces: a psychological and neural synthesis. *Trends in Cognitive Sciences, 6,* 261-266.

Pare, M., Richler, R. C., & Ten Hove, M. (2003). Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect. *Perception & Psychophysics, 65,* 553-567.

Radeau, M. & Colin, C. (2001). Object identity is not a condition but a result of intersensory integration: The case of audiovisual interactions. *Cahiers de Psychologie Cognitive-Current Psychology of Cognition, 20,* 349-357.

Rosenblum, L. D., Johnson, J. A., & Saldana, H. M. (1996). Point-light facial displays enhance comprehension of speech in noise. *Journal of Speech and Hearing Research, 39,* 1159-1170.

Rosenblum, L. D., Niehus, R. P., & Smith, N. M. (2007). Look who's talking: recognizing friends from visible articulation. *Perception, 36,* 157-159.

Rosenblum, L. D. & Saldana, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 318-331.

Rosenblum, L. D., Smith, N. M., Nichols, S. M., Hale, S., & Lee, J. (2006). Hearing a face: Cross-modal speaker matching using isolated visible speech. *Perception & Psychophysics, 68,* 84-93.

Schacter, D. L. & Church, B. A. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Schiller, P. H. & Carvey, C. E. (2005). The Hermann grid illusion revisited. *Perception, 34,* 1375-1397.

Schweinberger, S. R., Herholz, A., & Sommer, W. (1997). Recognizing famous voices: Influence of stimulus duration and different types of retrieval cues. *Journal of Speech, Language, and Hearing Research, 40,* 453-463.

Schweinberger, S. R., Herholz, A., & Stief, V. (1997). Auditory long-term memory: Repetition priming of voice recognition. *The Quarterly Journal of Experimental Psychology, 50A,* 498-517.

Schweinberger, S. R., Robertson, D., & Kaufmann, J. M. (2007). Hearing facial identities. *Quarterly Journal of Experimental Psychology, 60,* 1446-1456.

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics, 59,* 73-80.

Senkowski, D., Talsma, D., Grigutsch, M., Herrmann, C. S., & Woldorff, M. G. (2007). Good times for multisensory integration: Effects of the precision of temporal synchrony as revealed by gamma-band oscillations. *Neuropsychologia, 45,* 561-571.

Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive Brain Research, 14,* 147-152.

Shams, L., Kamitani, Y., Thompson, S., & Shimojo, S. (2001). Sound alters visual evoked potentials in humans. *NeuroReport, 12,* 3849-3852.

Sheffert, S. M. & Olson, E. (2004). Audiovisual speech facilitates voice learning. *Perception & Psychophysics, 66,* 352-362.

Shepard, R. N. (1964). Circularity in Judgments of Relative Pitch. *Journal of the Acoustical Society of America, 36,* 2346-&.

Soto-Faraco, S. & Alsius, A. (2007). Conscious access to the unisensory components of a cross-modal illusion. *NeuroReport, 18,* 347-350.

Soto-Faraco, S., Lyons, J., Gazzaniga, M., Spence, C., & Kingstone, A. (2002). The ventriloquist in motion: Illusory capture of dynamic information across sensory modalities. *Cognitive Brain Research, 14,* 139-146.

Stein, B. E. & Meredith, M. A. (1993). *The Merging of the Senses*. MIT Press.

Stein, B. E. & Wallace, M. T. (1996). Comparisons of cross-modality integration in midbrain and cortex. *Extrageniculostriate Mechanisms Underlying Visually-Guided Orientation Behavior, 112,* 289-299.

Sumby, W. H. & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustical Society of America, 26,* 212-215.

Summerfield, Q. (1992). Lipreading and Audiovisual Speech-Perception. *Philosophical Transactions Of The Royal Society Of London Series B-Biological Sciences, 335,* 71-78.

Suzuki, K. & Arashida, R. (1992). Geometrical Haptic Illusions Revisited - Haptic Illusions Compared with Visual Illusions. *Perception & Psychophysics, 52,* 329-335.

Thornton, I. M. & Kourtzi, Z. (2002). A matching advantage for dynamic human faces. *Perception, 31,* 113-132.

van Atteveldt, N. M., Formisano, E., Blomert, L., & Goebel, R. (2007). The effect of temporal asynchrony on the multisensory integration of letters and speech sounds. *Cerebral Cortex, 17,* 962-974.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences, 102,* 1181-1186.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia, 45,* 598-607.

VanLancker, D., Kreiman, J., & Emmorey, K. (1984). Recognition of famous voices forwards and backwards. *UCLA Working Papers in Phonetics, 59,* 114-119.

Vatakis, A., Navarra, J., Soto-Faraco, S., & Spence, C. (2007). Temporal recalibration during asynchronous audiovisual speech perception. *Experimental Brain Research, 181,* 173-181.

Vatakis, A., Navarra, J., Soto-Faraco, S., & Spence, C. (2008). Audiovisual temporal adaptation of speech: temporal order versus simultaneity judgments. *Experimental Brain Research, 185,* 521-529.

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A. L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience, 17,* 367-376.

Warren, D. H., Welch, R. B., & Mccarthy, T. J. (1981). The Role of Visual-Auditory Compellingness in the Ventriloquism Effect - Implications for Transitivity Among the Spatial Senses. *Perception & Psychophysics, 30,* 557-564.

Welch, R. B. & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88,* 638-667.

Zekveld, A. A., Kramer, S. E., Vlaming, M. S. M. G., & Houtgast, T. (2008). Audiovisual perception of speech in noise and masked written text. *Ear and Hearing, 29,* 99-111.

Zhou, Y. D. & Fuster, J. M. (1997). Neuronal activity of somatosensory cortex in a cross-modal (visuo-haptic) memory task. *Experimental Brain Research, 116,* 551-555.

Ehrenwörtliche Erklärung,

- dem Antragsteller, David Robertson, ist die geltende Promotionsordnung bekannt;

- der Antragsteller, David Robertson, versichert, dass er die Dissertation selbst angefertigt hat; er hat dabei die Hilfe eines Promotionsberaters nicht in Anspruch genommen, und alle von ihm benutzten Hilfsmittel und Quellen sind in seiner Arbeit angegeben;

- der Antragsteller, David Robertson, erwähnt im Gliederungspunkt „Acknowledgements“, welche Personen ihn bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskriptes unterstützt haben (entgeltlich/unentgeldlich);

- der Antragsteller, David Robertson, versichert, dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen;

- der Antragsteller, David Robertson, hat die Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht;

- der Antragsteller, David Robertson, hat die gleiche, eine in wesentlichen Teilen ähnliche oder eine andere Abhandlung bei einer anderen Hochschule bzw. anderen Fakultät nicht als Dissertation eingereicht;

Unterschrift _____ Ort, Datum _____

Lebenslauf

Name: David Robertson

Geburtsdatum: 24/12/1981

Geburtsort: Lanark, Schottland

Familienstand: Ledig

1992-1999	Peebles High School, Peebles.
1999-2001	Jura, University of Glasgow.
2001-2005	Psychologie, University of Glasgow.
2004	Nuffield Summer Student Researcher, University of Glasgow.
Juli, 2005	MA (Hons.) Psychologie, First-Class, University of Glasgow.
2005 – 2008	Wissenschaftliche Mitarbeiter, Institut für Psychologie, Friedrich-Schiller-Universität Jena.

Publikationen

Schweinberger, S. R., Robertson, D., & Kaufmann, J. M. (2007). Hearing facial identities. *Quarterly Journal of Experimental Psychology, 60,* 1446-1456.

Schweinberger, S.R., Casper, C., Hauthal, N., Kaufmann, J.M., Kawahara, H., Kloth, N., Robertson, D.M.C., Simpson, A.P., & Zäske, R. (2008). Auditory adaptation in voice perception. *Current Biology, 18,* 684-688.